

Evaluating the Assessments in an Orthopaedic Residency Program

by

© Nicholas Carl Smith MD

A Thesis submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Masters of Science

Faculty of Medicine

Memorial University of Newfoundland

May 2015

St. John's, Newfoundland and Labrador

Abstract

Introduction: Orthopaedic surgical education has undergone major change in the last 15 years. Work hour restrictions, public accountability, and government pressures have led to a paradigm shift in the execution of surgical training. The Royal College of Physicians and Surgeons of Canada (RCPSC) is adopting a competency-based training model in an attempt to ensure the quality of its future surgeons. **Objective:** To evaluate the reliability and validity of assessment methods of orthopaedic surgery residents as defined by the RCPSC's CanMEDs framework. **Methods:** A critical appraisal was undertaken that indicated a paucity of studies evaluating strategies for assessing surgical competencies in residency training programs. Staff surgeons assessed residents in day-to-day performance of duties using the Interprofessional Collaborator Assessment Rubric (ICAR) and the Surgical Encounters Form (SEF). The assessments were collected and measurements of percent agreement, Cronbach's alpha, and Fleiss Kappa were obtained. **Results:** For the ICAR percent agreement was 80.6 percent. Cronbach's Alpha measure averaged 0.662 and the mean Fleiss Kappa score was -0.218 (95% CI -0.400 to -0.089). For the SEF percent agreement was 90.9 percent. Cronbach's alpha averaged 0.865, 0.920, 0.934, 1.00 and 1.00 for the Medical expert, Technical skills, Communicator, Collaborator, and Advocate roles respectively. The mean Fleiss Kappa score was 0.147 (95% CI -0.071 to 0.364). **Conclusion:** Low inter-rater reliability results suggest low levels of assessor agreement and subsequently invalid assessment measures. Modification to assessment methods will be required before a valid competency-training program can be fully adopted.

Acknowledgements

The completion of this project has only been possible due to the knowledge, patience and understanding of several groups and individuals.

To Dr. Frank O'Dea and Dr. Heather Jackman. As my program directors during this endeavour you allowed me the freedom during a busy residency to complete this work. Without your understanding this would not have been possible.

To the staff and my fellow residents in the Department of Orthopaedics at Memorial University for your time and consideration during the completion of so many evaluations. You represent the core of this project and are the reason for the strength behind our training program.

To Dr. Yilmiz Yildaz for your expertise in statistics and your ability to guide me through rough waters. My frustration at times was overshadowed only by your patience.

To Dr Vernon Curran and Mark Hayward for your collaborative efforts with a surgical resident and your dedication to the improvement of medical education.

To my co-supervisor, Dr. Gerry Mugford and to my committee, Dr. John Harnett and Dr. Olga Heath. I could not have been blessed with a more experienced group for this journey. At every turn my questions were answered and my path was straightened.

To Dr. Andrew Furey, my co-supervisor and teacher. When asked if it was possible during residency you simply said, "Yes" and your faith in my ability has been the driving force for this project.

A thousand thanks and a lifetime of gratitude.

Dedication

This work is for my fiancé Anna, whose support, patience and understanding has given me the foundation from which to complete this project. Your own accomplishments during this time were inspirational and motivating.

And for my parents, who instilled in me a desire to learn and who gave me the tools to achieve all of my goals.

I love you all.

Table of Contents

	Page
Abstract	i
Acknowledgements	ii
Dedication	iii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
List of Abbreviations	viii
Chapter 1: Introduction	1
1.1 Background	1
1.2 Literature Review - Surgical	4
1.3 Literature Review – Non-surgical	13
1.4 Assessment Methods	18
1.5 Psychometrics of Assessment Tools	20
1.6 Analysis of Agreement	24
1.7 Purpose	26
1.8 Co-authorship Statement	28
Chapter 2: Collaborator Role	30
2.1 Introduction	30
2.2 Methods	30
2.3 Results	32
2.4 Discussion	34
Chapter 3: Medical Expert Role	37

3.1	Introduction	37
3.2	Methods	37
3.3	Results	38
3.4	Discussion	41
Chapter 4: An Issue with Statistics		44
Chapter 5: An Epidemiological Approach to the problems.....		49
5.1	Introduction	49
5.2	Methods	49
5.3	Results	51
5.4	Discussion	53
Chapter 6: Conclusion.....		57
Bibliography		60
Appendix A: Summary of surgical evaluation studies		72
Appendix B: Interprofessional Collaborator Assessment Rubric – Orthopaedic surgery		74
Appendix C: Health Research Ethics Board Approval for ICAR.....		82
Appendix D: Surgical Encounters Form.....		84
Appendix E: Health Research Ethics Board Approval for SEF		86
Appendix F: Abstract Collaborator Role		88
Appendix G: Abstracts Medical Expert Role		89

List of Tables

	Page
Table 1: Commonly accepted values of kappa	25
Table 2: Commonly accepted values of Cronbach's alpha.....	26
Table 3: Number of collaborator evaluations completed for each resident	32
Table 4: Combined Cronbach's alpha scores for Collaborator evaluations	33
Table 5: Cronbach's alpha scores for ICAR	33
Table 6: Percent agreement, Fleiss kappa scores and 95 percent confidence intervals for Collaborator role	34
Table 7: Number of surgical assessments completed for each resident.....	39
Table 8: Cronbach's alpha scores for Surgical Encounters Form.....	40
Table 9: Percent agreement, Fleiss Kappa scores and 95 percent confidence intervals for Surgical Encounters Form.	41
Table 10: Alternate measurements of inter-rater reliability.....	47
Table 11: Questions related to significant covariates	51
Table 12: Odds ratios, 95% confidence intervals and p values for SEF.....	52
Table 13: Covariates reaching significance	53
Table 14: Results of multivariate regression analysis.....	53

List of Figures

	Page
Figure 1: Literature search summary	4

List of Abbreviations

ACGME – Accreditation Council for Graduate Medical Education

CanMEDS – Canadian Medical Education Directives for Specialists

CIHC – Canadian Interprofessional Health Collaborative

GIOSAT – Generic Integrated Objective Structured Assessment Tool

HREA – Health Research Ethics Authority

ICAR – Interprofessional Collaborator Assessment Rubric

ITER – In-training Evaluation Report

OSATS – Objective Structured Assessment of Technical Skills

OSCE – Objective Structured Clinical Exam

PGY – Postgraduate year

RCPSC – Royal College of Physicians and Surgeons of Canada

SEF – Surgical Encounters Form

SPSS – Statistical Package for the Social Sciences

STSAF – Structured Skills Assessment Forms

Chapter 1: Introduction

1.1 Background

Surgical education has historically taken place under a mentorship model (Brieger, 1980). Physicians who had gained specialized medical knowledge would pass on their skills and training to the next generation, and on the mentor's subjective evaluation of a student's skills, they would graduate to the realm of the surgeon. This practice has been undergoing a fundamental change over the last several decades (Rose, 2009). With increased public demand for accountability (Canter, 2011), government pressure (Blum, 2011), advancing technology, work hour restrictions (Baskies, 2008) and financial limitations there has been a shift toward defined, objective, competency based learning and assessment. Aspiring residents are required to demonstrate a core set of knowledge and skills at an expected level before they can be allowed to practice without restriction.

The assessment of surgical residents requires a multidirectional approach. A surgeon has passed through several stages of training including medical school, residency, possibly subspecialty training and has committed to continuing professional development. At each of these phases he or she has many roles with different expectations of themselves, from the community and from their employers. Several questions arise: Who should perform surgical assessment? What are the expectations? What is the minimum standard? How does one assess technical skill? What format should be utilized? Which type of assessment method is best? What is the gold standard?

In 1996, the Royal College of Physicians and Surgeons of Canada (RCPSC) adopted the CanMEDS Physician Competency Framework as a “common set of essential abilities that all

physicians, regardless of specialty, need for optimal patient outcomes”(Frank, 2003). The seven components of the framework include: medical expert, communicator, collaborator, manager, health advocate, scholar and professional (CanMEDs, 2005). In 1999, the Accreditation Council for Graduate Medical Education (ACGME) launched a similar project in the United States. Identified core values include: patient care, medical knowledge, professionalism, interpersonal and communication skills, practice-based learning and improvement, and systems-based practice (Fitzgibbons, 2012). All medical training centers within Canada are required to align their programs with the CanMEDs framework and this system provides a guide for determining how to undertake surgical resident assessment. Each assessment point should be performed under the heading of one of the CanMEDs roles. Whether or not this is a valid approach to the new problems facing medical education has yet to be determined.

As the CanMEDs framework has evolved during its twenty years in use there has been a slow drift towards a competency-based curriculum. By the fall of 2014 all surgical training centers, and for that matter all medical education institutions within Canada, will be required to adhere to a competency-based format. Though this term is often confused and misused, at its core philosophy, this implies a structured pathway with regular, defined, objective measures of knowledge and skill with the final product being a well rounded, capable clinician (Parent, 2013). As surgical residents advance through the years of training their progress will be accelerated or slowed as deemed necessary by objective measures of their performance.

The Department of Orthopaedic Surgery at the University of Toronto is leading the way towards a competency-based surgical program (Ferguson, 2013). Given the issues of decreasing resident experience due to work-hour restrictions and modern patient safety needs, their department felt a need to explore major changes to their surgical curricula (Nauta, 2012). In

November 2013, they published their three-year experience with this new education model. Their “curriculum map” was designed such that the CanMEDs requirements were met throughout residency with regular assessments during the program. Their goals were to achieve competencies through modular based training, accelerate the pace of skills acquisition, diminish wasted time and evaluate residents frequently. They came to the conclusion that their model is a viable one with the potential to overcome some of the burdens facing medical education. Caution must be taken when interpreting these early results. Only fourteen residents had completed the competency program and enormous financial and manpower support was supplied to them through their department. This substantial support was not received by the residents completing the standard program and it may bias the results in favor of the heavily resourced competency based program. Though the University of Toronto orthopaedic residency program has developed new assessment methods for the CanMEDs roles, at great cost to the department, they have yet to demonstrate these as reliable or valid. This shift towards an objective, competency-based system will require the availability of reliable and valid assessment methods of surgical residents in all aspects of their training (Grantcharov, 2009). The demands of an objectively sound assessment system will require that programs produce or use stringently validated assessment methods, no matter what form they take.

1.2 Literature Review - Surgical

In June 2013 our original search was performed with an update in October 2014. Using the PubMed, EMBASE and Cochrane search engines the surgical literature was explored looking for validated surgical assessment tools based on the CanMEDs roles. All possible combinations of the terms [residents] + [CanMEDs] + [orthopedic] + [evaluation] + [surgery] and [education] were applied (Figure 1). RefWorks citation manager was used to organize the searches. Duplicates were removed. Relevant titles and abstracts were evaluated leaving 15 papers for inclusion. A further four papers were found during review of relevant bibliographies. The Canadian Orthopedic Association website was also examined for any relevant papers, talks and abstracts.

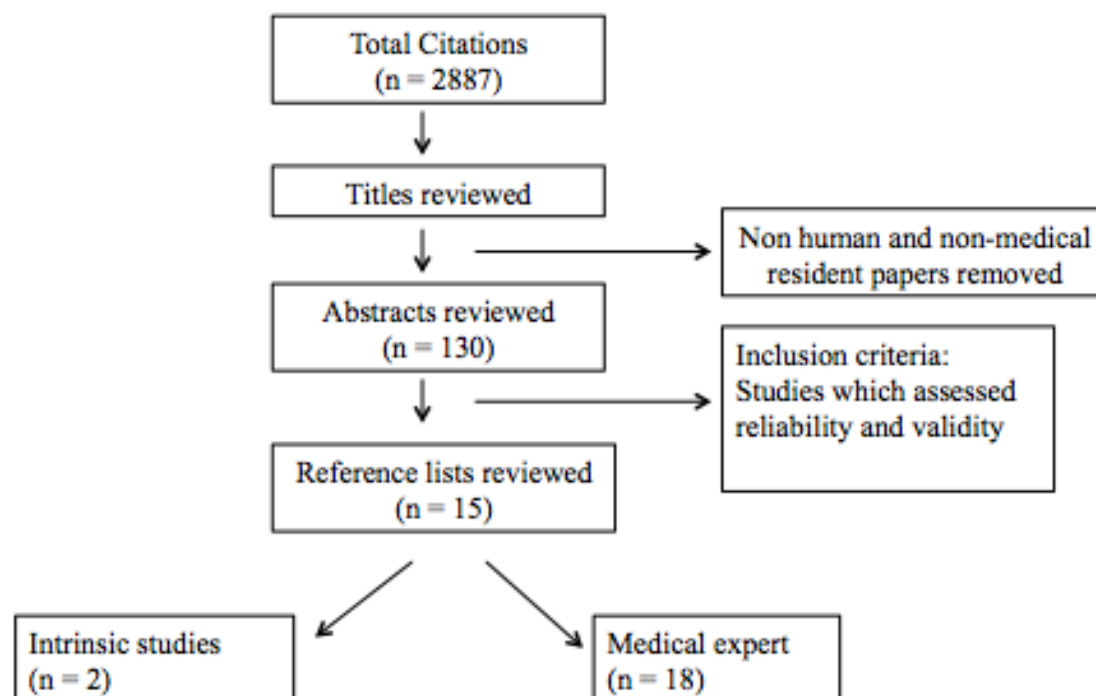


Figure 1: Literature Search Summary

The CanMEDs framework has given us a format on which to discuss the current status of assessment method literature in Canadian surgical training programs. The seven competencies can be broken down into two broader categories: the medical expert role and the non-medical expert roles, or intrinsic roles. The following review describes the available literature on the evaluation of surgical assessment methods.

The intrinsic roles are under-represented in the published literature. There are several reasons. First the medical expert role has historically been the major focus of surgical education. It has only been since the seven CanMEDs competencies came into existence that major emphasis has been placed on the non-medical expert roles. Secondly, they are more difficult to study. These roles are more difficult to objectively define and this has hampered enthusiasm to dedicate research endeavours in this field (Chou and Cole, 2008). Finally, from a surgeon's point of view these modern roles are often viewed as less important and more subjective (Arora, 2009).

Hanna et al (2012) examined the Manager role at McGill University. They used self-evaluation methods of senior surgical residents before and after a one-day course to assess the competency. Self-evaluations were the only methods performed and perceived improvement was noted. No measures of validity were assessed.

The most recent work has come from the University of Toronto, department of Orthopaedics. Dwyer et al (2014) created a six-station objective structured clinical examination (OSCE) to assess the six intrinsic CanMEDs roles. Twenty-five orthopaedic surgery residents performed the six-station examination. Validity was determined by comparing OSCE scores to in-training evaluation reports (ITER) completed over the previous twelve months and to an

ordinal ranking of resident performance created by the program directors. Reliability measures included Cronbach's alpha for inter-station reliability and an analysis of variance using training level as the independent variable and outcome score as the dependent variable. Interstation reliability measured 0.87 and there was a significant association between training year and examination scores. This group came to the conclusion that their OSCE was suitably reliable and valid for routine use in a surgical training program. No other studies evaluating the non-medical expert roles were found during our review.

As previously stated, the medical expert role has received more focus. A large volume of work has been dedicated towards creating modern assessment tools of the role that are both reliable and valid. Martin et al (1997), from the University of Toronto, performed some of the earliest work in this field, even before the medical role received its formal title. This study was designed using a six-station Objective Structured Assessment of Technical Skill (OSATS) exam evaluating a variety of surgical skills in both bench and live animal models. The purpose of the paper was to determine the reliability of the assessment tool while comparing live and bench models. Feasibility and practical application issues surround the tool had already been assessed and published by the group (Reznick, 1997). Twenty residents performed six stations on both bench models and live animals. Each was assessed with a task-specific checklist and a global rating scale. Reliability statistics included internal consistency and intraclass correlation coefficients. In the surgical literature the intraclass correlation coefficient is a common statistical representation of the agreement between multiple assessors. Cronbach's alpha ranged between 0.33 and 0.74. Intraclass correlations ranged between 0.64 and 0.72. Multivariate analysis of the tool demonstrated that training level alone was the only significant determinant of performance. This was felt to represent strong construct validity.

Closely related to Martin's work, Winckle et al (1994), also from the University of Toronto, recognized the need for developing new assessment tools for general surgery residents. They developed Structured Technical Skills Assessments Forms (STSAF) for three general surgical procedures: cholecystectomy, inguinal hernia repair and bowel resection. Each tool had a task-specific checklist and a global assessment scale. This cross-sectional cohort analysis was performed in 1994 and evaluated six junior and six senior residents performing forty-one operations, twenty-six of which had paired evaluators. Reliability measures included Cohen's Kappa and Pearson's correlation coefficients. Face validity for the STSAF was confirmed using experts prior to the study. Construct validity was evaluated with a Student's t-test comparing mean scores from junior to senior residents. Kappa values were 0.78 and 0.73 for the task specific checklist and the global rating scale. There was high correlation (0.89) between the two scores. Student t-test reached significance at <0.001 helping to demonstrate construct validity.

In 2004, Goff performed a multi-center cross-sectional analysis of a resident assessment tool created by the author and previously shown to be valid and reliable at a single institution (Goff, 2005). The purpose of this study was to assess the reliability and validity of the tool when administered across multiple gynaecology programs in the United States (Goff, 2002). This six station objective structured assessment of technical skills (OSATS) was administered to 116 residents from six training centers for a total of 696 evaluations between 2001 and 2002. Residents were scored with a task specific checklist, a global rating scale and an overall pass/fail judgement by three evaluators, at least one of whom had no previous experience with the resident. No significant difference was discovered comparing blinded to non-blinded judges. Reliability measures included internal consistency (Cronbach's alpha) and inter-rater reliability (intraclass correlation coefficients). Alpha ranged from .71 to .90. Intraclass correlation ranged

from .70 to .97. Construct validity was determined using one-way analysis of variance with the Student-Newman-Keuls test and residency year as the independent variable. They demonstrated that more senior residents had significantly better performance on all measures.

Roberson recognized the need for new assessment tools and set out to do so in a logical fashion (Roberson, 2005). The process must proceed in a particular order with: face validity, interobserver reliability, intraobserver reliability, construct validity and finally with confirmation of a pass standard. The purpose of their study was to test the reliability and in turn, the validity, of a tonsillectomy assessment tool, newly developed at Children's Hospital Boston. Between July 2002 and June 2004, a cross-sectional prospective cohort of 45 post-graduate year (PGY) three residents who performed tonsillectomies was carried out. Residents were assessed on both a task specific checklist and a global rating scale. Sixteen assessments had multiple evaluators for inter-rater observations making this a prospective evaluation. Percent agreement, defined as agreement within two points on a five-point scale, measured better than 97 percent. Weighted kappa scores (Fleiss Kappa) were calculated for every question on the tool. These ranged from negative values to greater than 0.90. The authors concluded that items with higher kappa would be more reliable for future tools. Construct validity was assessed in two ways. Resident's scores were compared to ten staff and fellow assessments. The resident scores were consistently lower than both staff and fellows. Secondly, residents who were assessed at multiple times during their three-month rotations trended towards significantly higher scores after their learning period.

Laeq (2010) performed a cross-sectional cohort study at Johns Hopkins Hospital. The purpose of the study was to evaluate the reliability, validity and feasibility of an assessment tool for endoscopic sinus surgery. The tool had previously been validated in the laboratory (Laeq, 2009) following a Delphi survey and this was the first attempt at clinical application. Eight

residents performed a single endoscopic sinus surgery and this was recorded on video. Five evaluators, all blinded to resident identity, assessed the video with the ability to fast-forward and rewind as deemed appropriate. The tool comprised of both a procedures checklist and a general rating scale. In total 40 assessments were performed. Cronbach's alpha demonstrated strong internal consistency at 0.85. Inter-rater reliability was shown with inter-class correlation coefficients and measured 0.62. Construct validity was assessed with a one-way analysis of variance to distinguish between resident training level. The tool provided a significant difference between junior and senior residents.

Ishman (2012) completed a cross-sectional cohort study at the Johns Hopkins School of Medicine Otolaryngology department. The purpose of the study was to evaluate the reliability of a two page OSATS assessment tool for paediatric laryngoscopy and rigid bronchoscopy. This evaluation tool had been created using the Delphi technique among experts and had been previously piloted by the same author (Ishman, 2010). This initial work was a non-blinded study that demonstrated good reliability and validity. The authors were concerned about the potential for confirmation bias in their initial study, and they subsequently designed the second evaluation in a blinded fashion to attempt to control for this bias. Fifty-two paired assessments were performed using both a task-specific checklist and a global rating scale. Faculty members were unfamiliar with the residents. Forty-five assessments had complete data sets available. Statistical measurements were done as both binary and continuous variables in order to evaluate reliability using Kappa and Intraclass correlation coefficients. Percent agreement ranged between 71.4 percent and 77.4 percent for binary variables. Kappa values were 0.38 to 0.54 for the binary assessment. Evaluation measures for continuous variables revealed a percent agreement between 42.9 percent and 71 percent, and intraclass correlation coefficients between 0.53 and 0.73. Alpha

ranged between 0.71 and 0.92. Only PGY two and three residents were assessed and therefore construct validity was not evaluated. Face validity had been confirmed in their 2010 study.

Laeq and Ishman would take their research further and evaluate the reliability and validity of a tonsillectomy evaluation tool (Ahmed, 2013). This included a task-specific checklist and a global rating scale that were created at their institutions. Eighty-three evaluations of their residents were completed and compared using average scores and Cronbach's alpha for internal consistency. This demonstrated high scores of 0.97 and was felt to satisfy construct validity as more operative experience led to significantly higher scores.

Moktar (2014), at the University of Toronto, developed a novel simulator for the assessment of casting techniques. They developed a video based assessment method of a casting simulation that yielded intraclass correlation coefficients of 0.88. They recognized the need to determine construct validity but had yet to do so. Golnik (2013) assessed an internationally created tool for the assessment of resident training in phacoemulsification (cataract surgery). Cronbach's alpha measured 0.92 as ten experts evaluated six recorded surgeries demonstrating internal consistency. Glarner (2013) developed an evaluation tool for laparoscopic colon resections that assessed both technical skills and the non-medical expert roles (termed NOTECH's in the study). They demonstrated face validity through staff agreement and construct validity by comparing resident scores through different years. There was a significant difference in the medical expert evaluations but not in the NOTECH scores. The reliability of an arthroscopic skills assessment tool was assessed by Koehler (2013). They demonstrated intraclass correlation coefficients of 0.83 for staff persons assessing recorded videos of a model knee arthroscopy. Other studies by: Benson (2012), Chou (2008), Larson (2005), Lentz (2001), Lin (2009), Palter (2012), Stack (2010), Johnson (1998), Jefferies (2007), Grober (2006), and

Mickelson (2008) have been performed that either required further work or were felt to be of less scientific merit

The currently available literature on evaluations of assessment methods has some notable shortcomings. The statistical methods used were confusing and inconsistent. All assessment tools were created using Likert scales of ordinal measure. Several studies presented statistics based on Likert scale data as if these were continuous variables. This leads to some potential for statistical bias. Martin (1997) performed the benchmark studies of resident assessment. Their models laid the groundwork for future papers that have been published in this field. Some of their statistical methods were flawed and this may have influenced several studies that would follow.

Blinding was another frequent concern that many authors noted. Winckel (1994) made the interesting observation that even if evaluators are blinded to the resident's skill level it may be difficult to blind completely. More senior residents tend to have increased skill and confidence. They also note that a highly structured task-specific checklist should help to limit this bias by leaving no room for individual interpretation. Each assessment item was either rated as either complete or incomplete with no in-between options. Skills were either demonstrated in full or not at all.

Roberson's (2005) study gives us perhaps the best model on which to base future work. By starting with a set of specific criteria that demonstrated face validity, they were able to reduce their tool to its reliable components. The product is a reliable, valid and feasible tool. By their own definition they need further studies to confirm the between-rater reliability and eventually pass criteria, but no other study had progressed this far. Of interest, no studies available to date

have evaluated currently employed assessment methods. Each was an attempt to create new and valid measures but no programs appear to have evaluated their own, tried and tested assessment regimes.

This wide range of assessment methods spans many of the surgical specialties. Most emphasis has been placed on creating valid assessment tools for individual surgical procedures. These methods will help to provide the backbone for a competency based system in the future. Understanding the breadth of evaluation options will reveal the scope of the void that will need to be filled.

1.3 Literature Review – Non-surgical

A complete examination of the assessment strategies of residents requires an evaluation of the non-surgical literature. The CanMEDs roles apply equally to training programs outside of the surgical setting. In October of 2014, a literature search was performed in the Pubmed, EMBASE and Cochrane Library engines looking for validated assessment methods of residents outside of surgical training programs. Combinations of the terms: [resident] + [CanMEDs] + [ACGME] + [evaluation] + [assessment] and [education] were applied to the search strategy. Studies examining the reliability and validity of resident and medical student assessment tools were included.

Busari (2014) performed a systematic review of the literature to determine if any reliable and valid assessment methods had been published for examining the ACGME and CanMED's system-based practice and manager roles, respectively. Their comprehensive review finished in November of 2012 but no validated assessment measures were identified. They recognized that while these roles have been established as important to the future practice of physicians, little attention has been paid to evaluation strategies of the roles themselves. They recommended future work be dedicated to establish the validity of assessment measures.

In the field of emergency medicine, Sherbino (2013) sought to determine the reliability of a clinical encounter card system for assessing medical students under the CanMEDs umbrella. The encounter cards require that the staff physician assess the student based on the medical expert role, up to two of other six roles, and overall performance. No training had been provided to the staff persons. They used a generalizability theory to determine their inter-item and inter-rater reliability. The scholar, collaborator, manager and health advocate roles were reported on

less than 25 percent of the assessments. For their instrument, 67 percent of the variability within the scores was related to the observer and not a student-based factor. They also noted that each of the CanMEDs ratings was highly correlated with the student's overall score. Though there was only speculation as to the cause of this finding, they raise a concern that each of the intrinsic roles may be too closely related to allow for raters to distinguish between them.

A second Canadian study from the emergency medicine field was completed in 2014 (Kassam). They sought to determine the reliability of a 24 question ITER modified for each specific year of training within their program at the University of Calgary. Their ITER data was collected from 2009 to 2011 and examined for reliability using Cronbach's alpha and for construct validity using an exploratory factor analysis with varimax rotation. The overall alpha score was 0.97 and the factor analysis revealed a five-factor solution that accounted for 79 percent of the variance. They came to the conclusion that their ITER demonstrated strong reliability with evidence of construct validity.

In the United States, the Academic Emergency Medicine consensus conference (Rodriguez, 2012) performed a systematic review of the literature to determine if reliable and valid tools existed for measuring the ACGME professionalism competency. They classified their findings into each of six headings: ethical knowledge and moral reasoning tests, direct observation assessment tools, survey-based assessment tools, critical incident reporting systems, portfolios and narratives, and simulated encounter observations. Though they identified several tools, none had been vigorously evaluated for validity, reliability, feasibility, educational impact or acceptability. Their consensus group made future recommendations for the development of validated assessment strategies of the professionalism role.

The department of Anaesthesia at the University of Ottawa developed a Generic Integrated Objective Structured Assessment Tool (GIOSAT) and sought to evaluate its reliability and validity (Neira, 2013). Their work focused on two videotaped, mock scenarios in a paediatric anaesthesia setting. This was then evaluated using the GIOSAT tool by four independent raters who were blinded to the residents training level. Reliability was measured with intra-class correlations for single raters and the average for four raters. Construct validity compared GIOSAT scores with residency training year. The average intra-class coefficient was 0.85 demonstrating strong reliability and there was a high correlation between GIOSAT scores and resident training level. Interestingly, when the tool was broken down into the seven CanMEDs roles there was strong correlation between training year and the medical expert role but not with the intrinsic roles.

In 2013, a child and adolescent psychiatry program evaluated a global rating scale designed to assess the six ACGME core competencies (Tomisato, 2013). Their initial three-staged design process allowed for modifications to the tool during practical application. In the final analysis they evaluated intra-class correlation with results ranging from 0.778 to 0.945 for the individual competencies. They came to the conclusion that their tool was reliable, valid and feasible. A relatively small sample size was the only limitation noted within the study.

As of 2011, Germany had developed a Medical Licensure Act that has been used to guide medical education in the country. Though not based on either the CanMEDs or ACGME guidelines it maintains several of their core principles. A questionnaire, translated simply to the “FKM,” has been developed to assess the reliability and validity of these competencies (Giesler, 2011). Six hundred ninety-eight medical students and 514 residents were surveyed between 2008 and 2011 using the 45-item assessment tool. Cronbach’s alpha was used to determine internal

consistency and remained consistent between 0.68 and 0.97. Construct validity was determined using t-tests and ANOVAs to assess differences between training level and competence level. Overall the questionnaire was felt to be reliable, valid and feasible.

ITERS have commonly been used during Rheumatology rotations as assessment measures but their reliability has been limited (Humphrey-Murto, 2009). In an attempt to improve the reliability of their assessment methods, a CanMEDs based evaluation form was developed for internal medicine residents rotating through the Rheumatology service. The University of Ottawa and McMaster University participated in the study. Residents were encouraged to have these forms completed daily during their one-month rotations. No formal training had been given to the evaluators. In total 637 assessments were completed for 73 residents. Reliability was determined through use of a generalizability coefficient. Each resident averaged 8.73 assessments during the rotation and 14 would be required to achieve a coefficient of 0.80. Eight of the 73 residents had completed an end of rotation OSCE. Numbers were low because of timing of the OSCE. Pearson correlation coefficient was used to compare results of the OSCE versus the ITER. At 0.48 the results were not correlated.

From January 2002 to December of 2004 an online, 360-degree assessment tool for competence was applied to psychiatry residents at the University of Washington. (Massagli, 2007). Nurses, allied health staff and medical students performed 935 evaluations of 56 residents over this period. These evaluations were performed at the end of each resident rotation. Cronbach's alpha revealed a reliability of 0.89. A reliability of greater than 0.8 could be achieved by only five nursing evaluations, compared to 23 ratings from medical students. More senior residents achieved high scores. The group felt this 360-degree evaluation tool was reliable, valid and feasible for the assessment of rehabilitation residents.

This literature review of the non-surgical studies provides similar conclusions to that of the surgical studies. The medical expert role is frequently correlated with resident training level and appears well described. Assessment methods of the intrinsic roles are less reliable. A general call for high quality studies to develop and confirm the reliability and validity of resident assessment methods spans both the surgical and non-surgical literature.

1.4 Assessment Methods

For hundreds of years the backbone of surgical assessment has been the essay style, written examination combined with direct preceptor observation. These served as the only assessment methods for young physicians. The quality of this strategy has been questioned for some time (Wanzel, 2002). First the content of written examinations and then the value of their results were evaluated (Brieger, 1980). This directly led to the creation of new formats including multiple-choice questions, extended matching options and clinical scenario pathways.

The next generation of assessment methods saw the introduction of the standardized oral exam, Objective Structured Clinical Exam and In-Training Evaluation Reports. These methods involve a one-on-one assessment by a preceptor with the goal of creating an evaluation scenario more aligned with clinical practice in both the cognitive and technical setting. Unfortunately these methods are often subjective and completed retrospectively which can bring their validity into question (Dent, 2009).

Given questionable validity with existing assessment methods there is a need to develop more psychometrically sound instruments. Computer simulations (Froelich, 2011), virtual reality, self assessments (Trajkovski, 2012), standardized patients (Hassett, 2006), objective structured assessments of technical skill (OSATS) (Chipman, 2009), point of observation assessments (Anderson, 2005), operative performance ratings (Williams, 2012) and interactive models (Moktar, 2014) have all been employed in recent years. Some areas, such as computer simulations, show promise (Koehler 2013) and some models are in widespread use, such as standardized patients (Ortwein 2011). They have several limiting factors including access, cost, complexity, lack of demonstrated efficacy and questionable applicability.

Another assessment method is the 360-degree model. In the medical field this implies feedback from nurses, physiotherapists, patients, peers and other allied health professionals as they interact with resident surgeons. This can potentially allow for a broader assessment of a resident physician's skills for both the medical expert and non-medical expert CanMEDs roles. Donnon (2014) recently performed a systematic review of the literature to determine the potential reliability, validity and feasibility of this method. Four studies discovered within the surgical literature support this as a potential option in terms of feasibility, applicability and outcome measures. Though not yet in widespread use, the 360-degree model may see broader applications in the future.

1.5 Psychometrics of Assessment Tools

Understanding the nature of the question being asked, and subsequently, the form the data will take is key to assessing the usefulness of assessment methods. The variable being examined can be either continuous or categorical (Hulley, 1988). Continuous variables can take any value along a numeric scale. Height, weight and age are examples here. Categorical variables can take only defined values as set out by the groupings. Some examples would be race, sex or smoking history. Categorical variables can be further subdivided into nominal and ordinal groupings. Nominal variables have no order. Again hair colour is an example. Ordinal measures, such as resident competency, occur in order from poor to good to excellent.

In order for a measurement to be valuable we need to trust the results. That is, any measured variable must be valid. Validity, attempts to discern if an assessment tool is actually measuring that which it purports to measure. Historically this involved the concepts of face, content, criterion and construct validity (Sackett, 1991). Face validity is simply a general determination of whether a tool measures that which it purports to measure by looking at its general parts. The item as a whole should appear to be a reasonable device for determining its outcome measure. This is a vague concept with only subjective analysis and no objective measures of strength. Similarly, content validity sets to determine if a tool appropriately includes all potential facets of a measurement scale. A tool measuring dietary intake must include a section for liquids as well as a section for solids. If either section is missing then the content is incomplete and conclusions drawn from this cannot be trusted. Criterion validity requires a comparison to a gold standard already supported in the system. Unfortunately in the creation of tools for assessment of surgical skills there is often no defined gold standard. Finally, construct validity seeks to determine if the conclusions being drawn are appropriate for the given tool. For

example, a measuring tape would have appropriate construct validity for measuring height, but not for determining weight. Unfortunately these terms are often confused in the literature and are so broad they do not lend towards clear, objective measurement.

A more modern process has arisen in recent years in order to clarify the individual components necessary to ensure validity. This can be broken down into five factors: content, response process, relations to other variables, consequences and internal structure (Cook, 2006). The content of a tool should represent the entire construct it is assessing. There should be no extraneous information or deviation from the spirit of the construct. On the other hand, there should be no missing, pertinent details. Secondly, the response process should demonstrate that a tool's outcomes reflect the user's thoughts during an assessment moment. Cook explains that if an evaluator, or a student, were to speak out-loud and describe their thoughts during an assessment, the tool should adequately reflect these vital moments. If there is the possibility to be good, bad, or ugly the response process must reflect this. In essence, a valid tool must be built on foundations that reflect in the mental process of the assessment. Next, any new assessment tool should be comparable to currently used methods and should most closely align with the gold standard. Similar assessment methods should correlate with each other. The fourth factor in determining validity is the concept of consequence. Does the score make a difference? Can we take some amount of meaning from the result of the measure and take action based on the results? Ideally any type of resident evaluation tool would aid in academic advancement, job applications, guidance towards extra training and identifying areas of weakness. Finally, the internal structure aspect of validity seeks to determine the reliability of the tool. Similar items with the tool should obtain similar results and these scores should be properly reflected by the measurement.

Reliability is the concept that a measurement tool can achieve reproducible results between users and at different points in time. Terms often used synonymously with reliability are repeatability, precision, accuracy and consistency. This concept is closely related to validity and is often considered as an integral component in determining validity. Errors in reliability can be either systematic or random and affect the validity of an instrument. Systematic errors occur in the same direction each time a measurement is performed. An example is a scale without proper calibration. Every time an object is weighed the result will be consistently incorrect by the amount of error in the calibration. This affects the accuracy of the measurement. Random errors occur differently for each evaluation. Scientists measuring heat loss from a system may experience random error if changes in wind temperature are not considered. In order to improve the reliability of evaluations several options have been proposed: standardize the measurement methods, train observers, refine instruments, automate instruments, and take repeat measurements (Hulley, 1988). In understanding this definition we see that reliability is a key component of validity. Reliability scores are necessary, but not sufficient, for determining the validity of a construct (Cook, 2006).

A concept separate from validity but no less important is feasibility. Daily evaluation tools should be inexpensive, easily accessible, applicable over a variety of situations, and easy to complete. This helps to ensure compliance from both the assessor and the trainee. Creating comprehensive assessment tools can pose a significant challenge given the large breadth of surgical education and the wide scope of potential future practice. Nonetheless, the ideal assessment tool should attempt to cover the entire scope of the knowledge and skills being taught while at the same time remaining practical enough to ensure its feasibility.

Finally an assessment tool must fulfill the criteria for accountability. It is within the changing scope of surgical education that new tools are being developed and a large portion of the driving force for this change is from external influence. The public and government agencies want proof that their trust and their funds are being directed towards a valuable, competent product. Assessment tools should be able to provide this reassurance.

1.6 Analysis of Agreement

If an assessment method proves to be valid this implies that the results are reliable. Taking this in reverse, if an instrument proved to be unreliable then it must also be invalid. An important component of reliability is a measure of the agreement between users. Agreement occurs in a variety of forms. Inter-rater agreement occurs between different users evaluating the same item. In the evaluation of assessment methods this implies different staff members assessing the same resident. Intra-rater reliability occurs as the same user evaluates an item at different points in time. Here this would be a staff person evaluating the same resident at multiple encounters (Cohen, 1960).

Measurements of this type are required to aid in the demonstration of valid assessments and several methods are frequently employed. The most simplistic measure of agreement is absolute percent agreement. In an evaluation study this is the number of times different staff persons agree on a resident's performance. Though easy to calculate it does not take chance into consideration (Cohen, 1960). If an evaluation method has only four potential responses there is a 25 percent chance that evaluators will agree on this alone. Though many studies of resident evaluation describe the percent agreement this is not substantial enough for validating assessment tools.

In order to deal with this, Cohen developed the kappa measurement. The formula reads: $K = \frac{p_o - p_c}{1 - p_c}$ where p_o is the observed proportion of agreement and p_c is the probability of chance agreement. The commonly accepted values of kappa are:

Table 1: Commonly accepted values of kappa (Landis, 1977)

Value of Kappa	Strength of Agreement
<0	Poor – Less than chance
0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.0	Almost perfect

The values range from 0 to 1 with higher values implying greater concordance between users. This formula functions well for nominal data but does not take into account the ordering of an assessment scheme. In other words, ordinal data requires a different consideration.

In order to utilize a kappa statistic to determine agreement within ordinal data a weighted kappa measurement should be used. Ordinal data implies that there remains some degree of agreement between users, even if there is not perfect agreement (McGinn, 2004). One rating of “excellent” is similar to a rating of “good” and a weighted kappa captures that similarity. Here the formula changes to: $K=1-qo/qc$ where q equals the disagreement of the measure. The possible values of kappa for a weighted value range from negative 1, implying agreement less than chance, to 1, implying perfect agreement (Kramer, 1981).

One final measurement of reliability that has been frequently discussed in the surgical evaluation literature is Cronbach’s alpha. This is a coefficient of internal consistency. In other words it asks the question “Do like items achieve like measures” (Tavakol, 2011)? Determining the average inter-correlation amongst the individual items achieves this. The standardized formula reads as: $N\hat{C}/v+(N-1)\hat{C}$ where N = numbers of items, \hat{C} = average inter-item covariance, and v = the average variance. The commonly accepted values of Cronbach’s are:

Table 2: Commonly accepted values of Cronbach's alpha

Cronbach's alpha (α)	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Appendix A is a summary table of the statistical methods used for surgical evaluation reliability studies to date.

1.7 Purpose

The purpose of this thesis is to evaluate the reliability of currently utilized surgical training assessment tools at one institution's orthopaedics residency program as they relate to the CanMEDs competencies. Though there has been a push towards the creation of modern reliable and valid assessment methods there is a lack of literature discussing the state of our current assessment methods.

The CanMEDs competencies have laid out a structural framework by which these assessment methods should be guided and when applied to a surgical training program they can generally be broken down into two broad categories:

- 1) Intrinsic roles
- 2) Medical expert roles

The purpose of this thesis is to evaluate the reliability of the Interprofessional Collaborator Assessment Rubric and the Surgical Encounters Form for orthopaedic surgery residents during routine assessment periods of the CanMEDs Collaborator role and Medical Expert role, respectively.

1.8 Co-authorship Statement

Evaluating the CanMEDs collaborator role in an orthopaedic surgery residency program

Principle Author

Nicholas Smith

Contributing Authors

Dr Vernon Curran provided input during the research design stage. He and his co-authors allowed us the use of the ICAR during our evaluation study. He was involved in the manuscript preparation.

Mark Hayward was involved in the design stage, practical research aspects and the data analysis. He was completing a parallel project in the department of Anaesthesia and our collaboration allowed problem solving and project completion.

Dr Andrew Furey provided the initial departmental support from the Orthopaedic surgeons. He was then involved in the research design and the manuscript preparation stages.

Evaluating the reliability of surgical assessment methods in an orthopaedic residency program

Principle Author

Nicholas Smith

Contributing Authors

Dr John Harnett was involved in the data analysis and manuscript preparation stages.

Dr Andrew Furey provided the initial departmental support from the Orthopaedic surgeons. He was then involved in the research design and the manuscript preparation stages.

Chapter 2: Collaborator Role

2.1 Introduction

At present, there is no tool utilized by our institution that evaluates a single specific section of the CanMEDs roles. In order to evaluate our ability to assess the intrinsic CanMEDs roles we had to first find and integrate a suitable tool into the orthopaedic surgery program. An Interprofessional Collaborator Assessment Rubric (ICAR – Appendix B) had recently been created through our Centre for Collaborative Health Professional Education in conjunction with the University of Toronto and the University of Ottawa (Curran, 2011). The evaluation characteristics of the ICAR closely resembled the competencies set out by the CanMEDs Collaborator role.

2.2 Methods

An orthopedic specific version of the ICAR was created at our institution. The original ICAR instrument was distributed to the program director, research coordinator and clerkship coordinator. They were asked to evaluate the tool for content validity, interpretability, ease of use and feasibility. Any section of the rubric that was felt to be unnecessary or irrelevant by two of the three evaluators was removed. Any comments they had on individual sections were noted and applied to the rubric if necessary. In total 25 competency questions were included from the original 31-question rubric. Each question could be answered on a four-point scale as either “minimal,” “developing,” “competent,” “mastery” or “not observable.”

The Health Research Ethics Authority granted full ethics approval for the study in March 2012 (Appendix C). Orthopaedic staff and residents were asked to participate in the study and

following a briefing session their written consent was obtained. In total six residents and ten staff surgeons out of a possible twelve participated in the study. All eligible residents participated. Only orthopaedics residents completing one of their core orthopaedics rotations were included. Over a period of six weeks, the staff orthopedic surgeons were asked to assess their residents based on the orthopedics ICAR during day-to-day clinical encounters. This included time in the clinics and in the operating room. The data was collected in unmarked, sealed envelopes that were distributed to each staff member on the day of the given assessment. The research coordinator coded the data as a third party. The primary researcher remained blinded to the staff and resident identities. The orthopedics residents were assigned a random two-digit number and the staff surgeons were assigned a random two-letter code. The research coordinator held the master key to the participant identities on a locked, password protected computer.

Data was collected and analyzed for internal consistency using Cronbach's alpha and for inter-rater reliability using percent agreement and Fleiss Kappa scores. The SPSS (Version 19 Copyright 2010) statistics program evaluated for internal consistency using Cronbach's alpha. Following this, the same data was entered into the AgreeStats2011 (Version 2 Copyright 2010) software to assess the percent agreement and Fleiss Kappa scores for weighted data.

The study design is a prospective single-blind cohort. In order to ensure that the criteria for a valid weighted kappa analysis were met 64 evaluations would need to be collected. Cicchetti stated that a proper kappa measurement required the total evaluations to be equal to the number of categories squared and multiplied by four (Cicchetti, 1977). The data collection period for this trial occurred daily over a six-week period.

2.3 Results

Ten staff members assessed a total of six residents during the six-week data collection period. One resident was assessed by only one staff and was therefore removed from the analysis. Each of the other five residents was assessed by at least two staff surgeons (resident 19 by two staff and all the rest by three staff members). Residents ranged from PGY two to PGY five. No first year residents were available during the course of this study.

Table 3: Number of collaborator evaluations completed for each resident

Staff	Residents					
	15	16	17	18	19	20
AB	1					
CA	1					1
DA	1			1	1	1
DB						1
AD			1			
BA			1			
BC			1			
CD				1		
AC					1	
CB		1		1		
Total	3	1	3	3	2	3

Test for internal consistency of evaluator ratings for residents revealed a mean Cronbach's Alpha of 0.662 (range 0.116 to 0.986).

Table 4: Combined Cronbach's alpha scores for Collaborator evaluations

Resident	Cronbach's Alpha
15	0.771
17	0.601
18	0.986
19	0.116
20	0.838
Mean	0.662

The ICAR contained six subheadings each of which represented one aspect of interprofessional collaboration. These included: communication, collaboration, roles and responsibilities, collaborative patient/client-family centered approach, team functioning and conflict management/resolution. Individual alpha scores were calculated for the subheadings within the ICAR. Most calculations were not possible because of a lack of variability with the rater responses.

Table 5: Cronbach's alpha scores for ICAR

Resident	Cronbach's Alpha					
	Comm	Coll	Roles	Client	Team	Conflict
15	NV	NV	NV	NV	NV	NV
20	NV	NV	NV	NV	NV	NV
17	NV	NV	NV	NV	NV	NV
18	NV	NV	NV	NV	NV	NV
19	0.571	NV	NV	NV	NV	NV

NV = No variability

A custom weighting scale was applied for the weighted analysis given the ordinal nature of the data. "Mastery" was rated 1, "competent" was rated 0.7, "developing" was rated 0.4 and "minimal" was rated 0.1. The weighted percentage agreement was 0.806. The mean Fleiss Kappa scores were -0.218 for weighted data.

Table 6: Percent agreement, Fleiss kappa scores and 95 percent confidence intervals for Collaborator role

Resident	Weighted Scores		
	% Agree	Fleiss	95% Confidence Interval
15	0.792	-0.293	-0.563 to -0.222
17	0.788	-0.172	-0.391 to -0.252
18	0.816	-0.293	-0.418 to -0.168
19	0.817	0.060	-0.172 to 0.29
20	0.816	-0.293	-0.456 to -0.091
Mean	0.806	-0.218	-0.400 to -0.089

2.4 Discussion

A literature review had revealed no studies which sought to determine the reliability, and in turn validity, of the intrinsic CanMEDs assessment methods. Gilbert (2010) and his team published the Pan-Canadian Collaborator Competencies through the Canadian Interprofessional Health Collaborative (CIHC). In turn, Curran and his interprofessional team created a rubric designed to measure the outcomes defined by the CHIC. This Interprofessional Collaborator Assessment Rubric could be adopted and used as a model for assessing the CanMEDS collaborator role. Given the many different measures of reliability used in the surgical education literature we elected to calculate those most commonly employed.

Our assessment tool represents an ordinal scale. Evaluations of agreement within an ordinal scale are not “all or none.” For example, if we consider two separate evaluations, a rating at the mastery level closely resembles a measure at the competent level. This would represent more agreement than two evaluations at the mastery and developing level. It is reasonable to assume that the difference between the “mastery” and a “competent” rating is less than the

difference between the “mastery” and a “developing” rating. For this reason weighted scales were used throughout the study.

The percent agreement is simply the number of times different raters agreed on the measurement. The weighted percent agreement for this study was eighty-one percent. This represents a fairly high value of agreement but does not take chance into consideration. With a four-point scale there is a 25 percent probability that evaluators will agree on chance alone. In order to assess the tool while taking chance into consideration we used weighted Fleiss Kappa scores. Our resident assessments consistently reproduced kappa values of less than 0 with a mean score of -0.218 (95 % CI -0.400 to -0.089). The highest value achieved was 0.06. This implies that the agreement between users was actually less than that predicted by chance alone. Residents were consistently rated either a “4” or a “3” on the 4 point scale with only one “2” and zero “1” used throughout the study. This phenomenon, in which raters assign the same rating for each point of assessment, is known as the ‘halo effect’ (Thorndike, 1920). With this in mind there was, for all practical purposes, a fifty percent chance that staff should agree on any given competency. Residents were either “4” or “3.” This led to the negative value of Kappa and implies that there is poor inter-rater reliability within our assessment tool.

A final measure of reliability utilized was Cronbach’s alpha. This tool seeks to determine if similar items within a matrix are resulting in similar outcomes. For example, if a questionnaire on food choices asks “Do you like fruit” and “Do you like apples” we would expect a relatively high level of concordance. Questions on vegetables would be more likely to result in different answers. If a tool has an acceptable level of homogeneity and is seeking to determine a single construct, than we would expect a high level of alpha. The average Cronbach’s alpha for this study was 0.662, demonstrating questionable internal consistency.

Given the low levels of a weighted kappa statistic it was determined that the reliability of the ICAR in an orthopaedic resident assessment setting is poor. This conclusion is somewhat conflicted given the high percent agreement and reasonable alpha scores. The compliance with the assessments was also poor. Over eighty forms were distributed with only 16 returned and only 15 acceptable for analysis. According to Cicchetti's formula we would require 64 evaluations to accurately determine agreement. Given the low weighted kappa value it is questionable if a study with a larger sample size would move the kappa to an acceptable level. However the poor response rate is certainly a weakness of this study.

Chapter 3: Medical Expert Role

3.1 Introduction

Surgical education requires not only a firm knowledge base but also a mastery of technical skills. This broadly falls under the Medical Expert role of CanMEDs. One of our program's assessment methods of surgical skill is done through the Surgical Encounters Form (SEF – Appendix D). This 15-item form incorporates four of the CanMEDs competencies: medical expert, communicator, collaborator, and advocate and well as a section for technical skill, in order to fully assess surgical competence.

3.2 Methods

Ethics approval was obtained from our Health Research Ethics Board (Appendix E). In July of 2013 individual meetings were held with the staff orthopaedic surgeons and the orthopaedic residents. During these sessions the purpose of the study was explained. Consent was obtained and all questions were answered. Staff surgeons were already familiar with the SEF. The three point grading scale was explained carefully. All comparisons were made to staff surgeons. A “3” is equivalent skill to that of a board certified surgeon, “2” is capable skill but not yet staff person ready and “1” is insufficient skill. A fourth category was available for “not observed.” The staff surgeons completed assessments during operating days for all orthopedics residents on service. Residents off service were excluded as well as off service residents covering the orthopedics team. Residents ranged from PGY one to PGY five. The staff and resident would agree upon a case for assessment during each operating day. An electronic copy of the form, which could be completed on hand held devices, was emailed to the evaluating staff person and they were encouraged to complete the form as soon as possible following the

operation. The form was submitted electronically to a third party (the program research coordinator). Upon completion of the study all assessments were coded such to keep the principle investigator blinded to the study results.

Data was collected and analyzed for internal consistency using Cronbach's alpha and for inter-rater reliability using percent agreement and Fleiss Kappa scores. The SPSS (Version 20 Copyright 2011) statistics program evaluated for internal consistency using Cronbach's alpha. Following this, the same data was entered into the AgreeStats2013 (Version 2 Copyright 2013) software to assess the percent agreement and Fleiss Kappa scores for weighted data.

Thirty-six evaluations were required in accordance with the research question assessing the inter-rater reliability of the tool through a weighted measurement of Fleiss Kappa. Cicchetti's (1977) method for ordinal data, where the number of categories squared and multiplied by four was employed for the original sample size calculation.

3.3 Results

Eleven staff members assessed nine residents over a six-month period. Eighty-eight assessments were collected. One contained no resident identification and was discarded, leaving 87 evaluations. Residents ranged from PGY one to PGY five.

Table 7: Number of surgical assessments completed for each resident

Staff	Residents									
	22	33	44	55	77	88	99	00	11	12
AA							3			3
BB										
CC										
DD								4		2
EE	2				6	2		1		
FF			4		1	3		1	2	
GG			1		5	4		1		
HH										
II	1		2			2		2	1	1
JJ	2	1	1		3	3			1	
KK		1	1		4	2				
LL					1			1		
MM				1						
NN			2			3		4		2
Total	5	2	11	1	20	19	3	14	4	8

Cronbach's Alpha measure averaged 0.865, 0.920, 0.934, 1.00 and 1.00 for the Medical expert, Technical skills, Communicator, Collaborator, and Advocate roles respectively.

Table 8: Cronbach's alpha scores for Surgical Encounters Form.

Resident	Cronbach's Alpha				
	Med Exp	Tech Skills	Comm	Coll	Adv
22	0.939	ID	1.00	ID	ID
44	0.757	1.00	NV	ID	ID
88	NV	0.818	0.915	1.00	ID
11	NV	1.00	ID	ID	ID
77	0.895	0.832	0.822	1.00	1.00
99	NV	NV	ID	ID	ID
00	0.909	0.960	1.00	1.00	1.00
12	0.823	0.912	ID	1.00	1.00
Mean	0.865	0.920	0.934	1.00	1.00

Resident 33 had only two evaluations performed and was insufficient for Cronbach's analysis. NV = No variability, ID = Insufficient data.

The Agreestats2013 linear weighting scale was applied. The average weighted percentage agreement was 0.909. The mean Fleiss Kappa score was 0.147 (95% CI -0.071 to 0.364) for weighted data.

Table 9: Percent agreement, Fleiss Kappa scores and 95 percent confidence intervals for Surgical Encounters Form.

Resident	Weighted Scores		
	% Agree	Fleiss	95% CI
22	0.674	0.188	-0.27 to 0.645
44	0.916	0.433	0.073 to 0.792
33	0.714	0.142	-0.627 to 0.913
88	0.902	0.222	0.082 to 0.361
11	0.948	0.111	-0.074 to 0.297
77	0.818	0.022	-0.05 to 0.095
99	0.939	-0.304	-0.614 to 0.007
00	0.468	-0.097	-0.195 to 0.002
12	0.841	-0.095	-0.189 to -0.002
Mean	0.909	0.147	-0.071 to 0.364

Resident 33 had only 2 evaluations performed therefore Fleiss = Cohen's Kappa for 2 raters.

3.4 Discussion

Before starting the second project we sought to ameliorate the shortcomings of the Collaborator evaluation study. On discussion with the staff and residents, the feasibility of the Collaborator role assessment needed improvement. The paper form was long and cumbersome with overly complex wording. Hulley and Cummings proposed five key steps to improve reliability measures of assessments: standardize the measurement methods, train observers, refine instruments, automate instruments, and take repeat measurements (Hulley, 1988). Staff members were trained on the correct definitions and uses of the Surgical Encounters Form. Strict definitions of each category were employed. Complex wording of categories was simplified and shortened. We created an online version of the instrument that could be completed on mobile devices immediately after the observed procedure, and in doing so removed some of the potential for losing assessments and for recall bias. The new electronic form was

emailed to the staff each day they worked with a resident and took approximately two-to-three minutes to complete. Finally we performed two six-week assessment periods so as to improve our total number of responses. Our literature review yielded several studies that explored novel evaluation methods for surgical skills acquisition, but none that examined currently utilized methods.

During evaluation of the surgical encounters assessments eighty-eight forms were completed with 87 being suitable for analysis. The weighted percent agreement was 91 percent, which supports a high inter-rater agreement but did not take chance into consideration. Weighted kappa scores were 0.147 (95% CI -0.071 to 0.364). This demonstrates slight agreement between users when chance is considered. The scores here were significantly higher compared to the Collaborator study but were still lower than expected.

Cronbach's alpha was assessed. The SEF differs from the ICAR in that it incorporates several of the CanMEDs competencies, not just the Medical Expert role, as staff members assess surgical competence. In order to ensure heterogeneity did not falsely affect the results, each of the separate competencies alpha scores was determined for each of the roles individually (Table 8). Significant numbers of missing ("not observed") values and data that had no variability made some scores unattainable. The average alpha score for the Medical Expert role was 0.865 and for the Technical Skills role was 0.920. The Communicator, Collaborator and Advocate roles had values of 0.934, 1.00 and 1.00 respectively. An alpha of 0.865 represents almost perfect agreement. Caution must be taken in analyzing the final four alpha values. Such high scores likely represent a lack of variability within the tool, and though concordant, may not be reliable.

The reliability of the Surgical Encounters Form is questionable. High values of percent-agreement and Cronbach's alpha would seem to support its reliability but low weighted kappa scores suggest a less robust measure.

Chapter 4: An Issue with Statistics

The purpose of the study was to determine one aspect of validity using statistical measures of reliability. Aspects of validity can be difficult to demonstrate in concrete terms but reliability lends itself to an objective measurement. Our literature search revealed one of the fundamental concerns when performing an evaluation of assessment methods; what is the gold standard measurement of reliability? In an attempt to demonstrate reliability authors have utilized percent agreement, kappa, alpha, concordance and absolute score improvements.

We elected to calculate three of the most commonly used measures of reliability as seen in the surgical evaluation literature, recognizing them for both their strengths and weaknesses. Weighted percent agreement measures were quite high within our study. Unfortunately this overly simplistic measure of reliability does not account for chance. Early on we recognized that staff persons used a narrow range of values on an already narrow scale. The Collaborator assessment used a four-point likert scale but over the fifteen completed forms only one “2” and not a single “1” was selected. This may imply that the residents were all performing at or above average, that the staff persons were unclear of the distinctions, or it represents an unwillingness to rate a resident at the lower end of the scale. We helped to address this issue during the surgical encounters study through our teaching session with the staff members. This study contained a three-point likert scale with each of the categories clearly defined. Here the gold standard comparison became “staff person competent skill.” With a standard by which to compare each encounter, the full use of the scale was achieved.

Cronbach’s alpha was our second tool of choice. Conceptually this measurement fits well into evaluation literature and has often been used. Unfortunately there are several issues. First,

alpha does not handle missing data well. In medical literature there is often a category “Not assessed,” as was present in both of our studies, and this must be treated as an incomplete data set. In reality the data was not “missing” but the formula for alpha is not well able to deal with this. Secondly, heterogeneity within a tool must be carefully dealt with. The goal of alpha is to demonstrate the probability that similar assessment points come to similar conclusions. A tool measuring surgical skill has items related to technical merit and others related to interprofessional conduct. In other words the assessment tools are heterogeneous. This leads to difficulty generating a single alpha value for a given tool that has a built in assumption of unidimensionality. During the Collaborator study we were able to calculate a total alpha score. Individual scores for each of the six sub-headings were not possible because of a lack of variability within the responses. The total score may be skewed by the heterogeneity of the study as a whole. Each of the sub-headings assesses a unique aspect of collaboration and it may not be possible to combine these results. A larger volume of assessments may remedy this problem.

The surgical encounters study made improvements on the alpha measures. With little missing data and the ability to break each of the individual roles down separately we saw a score of 0.865 for the medical expert role. Low and high numbers of assessments can artificially inflate and deflate the scores. Small sample sizes will have artificially low alpha values while larger tests will create larger values of alpha. Some authors have suggested a maximum value of alpha to be 0.90, and any value above may reflect an erroneous result (Schmitt, 1996). Sijtsma (2009) discusses these problems and demonstrates in mathematical terms the results of a reliance on alpha. He goes so far as to say that Cronbach’s alpha is not a true measure of internal consistency and may have little to no value in the assessment of agreement. For all of these reasons we must accept our intrinsic alpha measures with caution.

The mean weighted kappa scores achieved through our evaluation studies were -0.218 (95 % CI -0.400 to -0.089) for the ICAR and 0.147 (95% CI -0.071 to 0.364) for the SEF respectively. These represent low to poor inter-rater agreement. Though kappa seems to be the most robust of our statistical measures, there are still concerns with its use. Though the kappa measurement has dominated the surgical literature, it has well known paradoxes that bring its value into question (Feinstein, 1990). If the probability of chance agreement among raters is high then the correction process can convert relatively high-observed percent agreement scores into low kappa values. Unfortunately assessment scales are often done on three, four or five point likert scales with relatively high likelihood for chance agreement. This may be a concern for the Collaborator study with a 25 percent chance agreement and for the Surgical Encounters study with a 33 percent chance agreement. If the paradoxes of alpha could be avoided it may be that there is reasonable agreement within our tools as reflected in the percent agreement.

Gwet elaborates on these concerns in his “Handbook on Inter-rater Reliability,” and suggests that more modern measures of agreement may be more suitable (Gwet, 2012). Though beyond the scope of this thesis, alternate formulas used to assess agreement, such as Gwet’s AC1 and Brennan-Prediger, may provide a more reliable and statistically stringent value of inter-rater agreement. Using the AgreeStats2013 software we were able to calculate these values for the Surgical Encounters Form data.

Table 10: Alternate measurements of inter-rater reliability

Resident	Weighted			
	Gwet	95% CI	BP	95% CI
22	0.313	-0.096 to 0.722	-0.266	-0.131 to 0.662
44	0.696	-0.368 to 1	0.505	0.179 to 0.832
33	0.604	-0.001 to 1	0.429	-0.323 to 1
88	0.855	0.768 to 0.942	0.734	0.622 to 0.846
11	0.939	0.870 to 1	0.861	0.727 to 0.994
77	0.643	0.570 to 0.715	0.508	0.44 to 0.576
99	0.936	0.612 to 1	0.879	0.501 to 1
00	-0.061	-0.072 to -0.05	-0.064	-0.073 to -0.55
12	0.738	0.651 to 0.825	0.571	0.474 to 0.668
Mean	0.697	0.399 to 0.867	0.486	0.385 to .763

Gwet = Gwet's AC1, BP = Brennan-Prediger, 95% CI = 95 percent confidence interval.

Resident 33 had only 2 evaluations performed.

Though the confidence intervals remain wide, which may imply an underpowered study, a glance at these newer values seems promising. A mean weighted agreement score of 0.697 seems to be more in keeping with the attitudes of the surgeons towards the resident assessments. Though these measures require more stringent assessment, they may provide an early look at the future of inter-rater reliability measures.

An option may be to consider interclass correlation measures such as Kendall's tau or Spearman's rho. These statistics work to determine if there is correlation between two sets of data. There are several options for setups within the resident evaluation studies. Each staff person assessment of a resident provides one set of data. Comparing all possible combinations of data sets and calculating this mean would provide the most powerful estimates of correlation. Banerjee (1999) describes a method of calculating correlation estimates for more than two raters, but the properties of this calculation are not fully understood and this method is not widely used.

A second option would be to compare groups of residents according to training year. By doing similar pairwise correlation measures of junior versus senior residents we would be able to determine a tool's ability to differentiate between training year. This format would be a less powerful calculation because of the broad definitions of junior and senior residents.

Chapter 5: An Epidemiological Approach to the problems

5.1 Introduction

After completion of the two projects yielded no definitive answers, we sought to approach the problem of evaluating the assessments in a different manner. Martin et al (1997) had previously discussed that the level of resident training was the only reliable predictor of score during their OSATS evaluations. Goff (2002) had also demonstrated reliability of their tool by confirming that senior residents achieved higher scores on their evaluations. The most methodologically sound papers consistently demonstrated the ability to distinguish level of training as a major factor in reliability analysis. The purpose of this new assessment was to objectively evaluate the Surgical Encounters Form's ability to differentiate between skilled and unskilled orthopaedics residents in terms of the CanMEDs competencies. A tool's ability to distinguish between junior and senior residents reflects on its validity. A secondary objective was to create a CanMEDs based, epidemiological model for the prediction of successful procurement of staff person equivalent surgical skill.

5.2 Methods

An updated literature search was performed in February 2014 on the PubMed, EMBASE and Cochrane search engines using the terms [resident] + [evaluation] + [competence] + [epidemiology] and [surgery]. No study had objectively evaluated assessment methods of orthopaedic surgical skill or previously attempted to apply an epidemiological approach to resident assessment.

Using the data acquired from the Surgical Encounters Form, residents were divided into two categories. Junior residents were defined as PGY 1 to PGY 3. The maximum amount of time on service at the beginning of the study for a PGY 3 was 10 months. Senior residents were defined as PGY 4 to PGY 5. The minimum amount of time on service to this point for a PGY 4 was 22 months. We defined less than or equal to “2” on the SEF implying not yet staff surgeon quality and “3” being equal quality to that of a staff surgeon.

Each of the 15 questions within the SEF was individually analyzed as a separate covariate as they related to the CanMEDs roles. To measure the association between resident level and rating on the SEF, odds ratios with their corresponding confidence intervals were estimated, and significance of the tests were obtained by p-values using a Chi Squared distribution and 1 degree of freedom. Sample size calculations were performed for each of the significant covariates defined at 0.05 level of significance and a power of eighty percent (Hsieh, 1998). Microsoft excel was used for the initial statistical calculations (Microsoft excel 2013 version 15.0.45). Each variable which was significantly different between junior and senior residents, as defined by $p < 0.05$, was extracted and inserted into a multiple logistic regression model (Table 11) (“R” stats 2013 version 3.0.3). The exposure, or explanatory variable, was the time spent within the orthopaedic training program. Junior was defined as PGY three or less and senior was defined as PGY 4 or higher. The outcome, or response variable, was the rating on the Surgeical Encounters Form between 1 and 3. The initial model was constructed using all significant variables. CRAN R statistical software was used to fit multivariable logistic regression models. Specifically, the “glm” function in R was used. As each model was fitted, the least significant variable was extracted. All combinations were assessed.

Table 11: Questions related to significant covariates

Covariates
1.b) Understands risks involved in surgery
2.c) Competent in surgical approach and dissection
2.f) Competent and safe throughout procedure

5.3 Results

Each of the covariates fell under one of five headings: medical expert, technical skills, communicator, collaborator, and advocate. The odds ratios were estimated to compare the odds of a “3” rating in the group of senior residents with the odds of a “3” rating in the group of junior residents for each covariate. These ranged from 0.7 to 3.9 with the average confidence interval being 0.43 to 3.97 (Table 12). Questions 1b, 2c and 2f were each significant with $p < 0.01$, 0.02 and 0.01 respectively, and fell under the medical expert and technical skills headings (Table 13). We also included question 1a as a significant covariate, with a p-value of 0.059, to be included in the multivariable logistic regression models. Though this did not strictly meet the criteria for significance, given the small sample size and its trend towards 0.05 we included it in the regression analysis. None of the questions under the communicator, collaborator or advocate headings reached significance. We used the “glm” function in R to fit the multivariable logistic regression models. No combinations of variables demonstrated significance (Table 14). The interaction between covariate 1b and 2c had the lowest p-value but did not reach significance at $p = 0.23$. In order to determine significance within 0.05 at eighty percent power, sample sizes were determined to be 197, 121, 182 and 158 for questions 1a, 1b, 2c and 2f, respectively.

Table 12: Odds ratios, 95% confidence intervals and p values for SEF

Question	Level	Rate 3	Rate <3	Odds Ratio	95% CI	p-value
1a	Senior	10	20	2.67	0.541 to 4.33	0.0598
	Junior	9	48			
1b	Senior	13	17	3.20	0.624 to 4.40	0.0171
	Junior	11	46			
1c	Senior	9	20	2.35	0.501 to 4.19	0.109
	Junior	9	47			
2a	Senior	9	13	2.34	0.453 to 4.61	0.147
	Junior	8	27			
2b	Senior	10	20	2.25	0.511 to 3.95	0.116
	Junior	10	45			
2c	Senior	12	18	3.26	0.600 to 4.65	0.0205
	Junior	9	44			
2d	Senior	10	20	1.75	0.472 to 3.45	0.267
	Junior	12	42			
2e	Senior	8	14	2.29	0.387 to 5.30	0.211
	Junior	5	20			
2f	Senior	10	20	3.92	0.579 to 5.65	0.0146
	Junior	6	47			
3a	Senior	8	19	0.64	0.301 to 2.26	0.389
	Junior	19	29			
3b	Senior	8	15	0.70	0.292 to 2.51	0.515
	Junior	16	21			
4a	Senior	7	18	0.80	0.316 to 2.62	0.683
	Junior	16	33			
4b	Senior	7	9	0.92	0.270 to 3.44	0.897
	Junior	11	13			
5a	Senior	7	13	1.08	0.296 to 3.61	0.908
	Junior	8	16			
5b	Senior	7	9	1.56	0.317 to 4.63	0.518
	Junior	7	14			

Table 13: Covariates reaching significance

Covariate (Question)	OR	95% CI	p-value
1.a) Understands indications for surgery	2.66	0.54, 4.33	0.060
1.b) Understands risks involved in surgery	3.19	0.62, 4.40	0.017
2.c) Competent in surgical approach and dissection	3.25	0.60, 4.64	0.020
2.f) Competent and safe throughout procedure	3.91	0.57, 5.65	0.015

Table 14: Results of multivariate regression analysis

	Covariate	OR	95% CI	p-value
4-Covariates	1.a	1.04	0.18, 6.07	0.965
	1.b	0.60	0.10, 3.64	0.569
	2.c	0.55	0.14, 2.23	0.402
	2.f	0.58	0.14, 2.38	0.447
3-Covariates	1.b	0.61	0.17, 2.19	0.445
	2.c	0.55	0.14, 2.23	0.402
	2.f	0.58	0.14, 2.38	0.448
2-Covariates	1.b	0.44	0.13, 1.57	0.311
	2.c	0.53	0.16, 1.80	0.208

5.4 Discussion

The current project sought to objectively determine if the SEF could differentiate between junior and senior residents in terms of their CanMEDs role competencies. Senior residents perform better than junior residents in their medical knowledge and their technical skill, but the CanMEDs platform demands that a well-rounded physician be trained in other areas (CanMEDs, 2005). Communication, collaboration and advocacy are three of the other roles assessed within the Surgical Encounters Form. The results of this study should be considered in two distinct groups. First, nine questions (covariates) fell under the medical expert and technical skills headings with three of them reaching significance and a fourth approaching significance at

$p = 0.059$. In the simplest interpretation of this data, we would state that senior residents are more knowledgeable and skilful than junior residents. This appears to be accurate and is supported by the SEF. Odds ratio estimates for these nine covariates ranged from 1.75 to 3.92. The odds of having staff person equivalent skill are 1.75 to 3.92 times greater in the senior than the junior residents.

The second group of variables to consider are the non-medical expert CanMEDs roles: communicator, collaborator and advocate. These did not demonstrate significance for any of the six covariates and did not seem to trend towards significance. Their p-values were higher than any of the medical expert roles, ranging from 0.39 to 0.90. The odds ratios were also significantly lower ranging from 0.64 to 1.55. The non-medical expert covariates were not able to predict junior versus senior residents.

It seems the SEF is able to predict which residents are more senior in terms of their surgical skill and medical knowledge. Though only four results reached significance, it appears the tool performs well in this regard. If a senior resident was consistently failing to achieve scores of 3 on the SEF we have objective evidence to say they are not yet competent in a given skill. The SEF is valid in this regard. On the other hand if a junior resident consistently produced scores of 3 he or she may be eligible for more expedited skills acquisition. It is granted that observer bias may be present due to a lack of blinding of the assessments. The non-medical expert roles results are less clear. The tool does not differentiate between junior and senior residents in these competencies. Without there being a significant difference here it is not possible to use the SEF, in its current form, to assess residents in terms of the intrinsic CanMEDs roles.

After determining the significant covariates we proceeded to create a model for predicting resident success. Are there certain elements of training that could prove to be more predictive of success or failure when analyzed together? Could we use this method as an adjunct measure of reliability? It is likely that there exist interactions between assessment items that can help to predict success or failure in the acquisition of surgical skill. By taking this epidemiologic approach to resident training assessment we can use logistic regression models to further define our covariates. The medical expert roles and technical skills roles of the SEF demonstrated significance and are closely related to each other. Each of these significant items was inserted into the model as described. Unfortunately, for this study no significant interactions were found.

Several interpretations can be made here. First, the study may be underpowered to find these differences. This is supported with the wide confidence intervals. We estimated our mean required sample sizes to be 165 for this particular series. Secondly, the residents may be no different in these skill sets. Either they enter the program already at a fully trained level, or the program fails to confer these new skill sets onto the residents. We may have watered down the results by considering the third year residents as juniors. Their skill sets may be at a junior or senior level depending on the situation. For this study, even if we remove them from analysis, there was no significant difference between the junior and senior residents for the non-medical expert roles. Finally, the tool may be invalid for assessing these roles. In that regard we return to the discussion on validating the tool for the purpose of resident assessment.

Though not directly measuring the inter-rater reliability of our tool, this approach provides a novel circumvention of the statistical problems encountered. If a tool can differentiate between junior and senior residents over many assessments then it demonstrates reliability. Unfortunately, elements of the SEF do not achieve this goal. The intrinsic roles are

not reliably measured. This seems to be in keeping with the attitudes of the surgeons performing these evaluations (Hopmans, 2013). For thousands of years surgical skills has been learned through a mentorship model and the educators in this field have developed sound and reliable teaching methods for this skills acquisition. The “newer” competencies such as communicator and collaborator have not historically been within the realm of the surgeon. Our epidemiological evaluation of the form seems to support that fact.

Chapter 6: Conclusion

The Royal College of Physicians and Surgeons of Canada has become a world leader in medical education through the CanMEDs initiative. Countries and programs around the world look to Canadian models for direction and guidance. As such, the RCPSC has both a national and international responsibility to rigorously assess its model of medical education. This truth comes to the forefront with the upcoming changes to the CanMEDs roles and the push towards a purely competency-based education system. One important aspect of this upcoming change is the ability to validly assess the training progress of medical students and residents alike. An in depth evaluation of the literature revealed a dearth of support for currently utilized assessment methods and a lack of methodologically sound studies for the guidance of future work. The purpose of this project was to evaluate the reliability of two currently employed assessment methods within one orthopaedic surgery residency program, and in turn comment on the validity of these methods.

As discussed, validity is a fluid concept that has undergone changes in the last several years. Cook (2006) provided the emerging criteria from which a surgical program can be evaluated. The content, response process, relations to other variables and consequence are all criteria that are difficult to objectively define during a short evaluation period. These factors must be constantly assessed and reassessed and adapted based on the needs of the program, the culture of the time and the requirements of external sources. Therefore, in order to comment on validity we elected to assess the internal structure, or reliability, of our assessment methods. Without reliability an assessment tool cannot be valid. Reliable scores are necessary, but not sufficient, for the demonstration of validity.

The determination of reliability is a key component in producing valid assessment tools. Other areas of surgical practice have demanded similar scrutiny. Furey (2004) sought to determine the reliability of commonly utilized fracture classification systems in an orthopaedic setting. The orthopaedic literature has been flooded with classification tools for the purpose of determining prognosis and directing treatment. Without sufficient reliability these tools would be invalid and any action taken based on the classification of fracture could be potentially harmful. They determined that for three commonly utilized fracture classifications there was low to moderate inter-rater reliability.

Though there was disagreement between the different statistical options used, it appears the overall reliability of the CanMEDs assessment methods of orthopaedics residents was low to poor at our institution. By implementing changes between the studies based on Hulley and Cummings (1988) suggestions for improving reliability scores we were able to improve our results. This was seen with the stronger reliability measures in the Surgical Encounters Form. Though this is encouraging, there is room for more improvement than has already been demonstrated. Ensuring the validity of future tools will require compliance and motivation from staff persons and residents alike.

We have demonstrated that as the feasibility of an assessment tool improves the compliance of staff surgeons and residents alike greatly improves. Simple modifications between our two assessments greatly increased the number of assessments completed, improved the surgeon's perceptions of the importance of the assessments and saved the staff persons considerable time and effort. The importance of feasible tools cannot be understated. Accountability must come at the university, provincial and national levels. Without combined efforts to regulate our teaching and assessment methods the quality of the surgeons we create

will suffer (Kamath, 2011). Radical changes to the CanMEDs implementation of medical education are likely to cause significant stress and in turn adaptation from each of these bodies. This study is a local attempt to hold us accountable for our education and assessment methods, and to stand by the medical students, residents and eventual staff surgeons we create. Having uncovered both positives and negatives within the assessment methods of our residents we must use this knowledge to make meaningful adjustments. In a period of significant change at the national level, it will be the responsibility of each individual program to closely monitor the effects both of these forces will have on resident education. Our group would recommend frequent evaluations of assessment methods and maintenance of effective communication between residents, staff, program administrators and the RCPSC. The coming years will have a major impact on the direction of surgical education both in Canada and internationally. Without a close monitoring, as suggested by our results, there is the potential for unintended, negative consequences.

A lack of scientific support for the assessment methods of surgical residents must be overcome in order to satisfy the increasing demands of the changing medical education system and of the taxpayers who support this system. With the principles of a valid assessment laid out, and the statistical options made clearer, this thesis attempts to provide a model for future analysis of surgical education. Problems exist within our education curriculum and by recognizing these we have taken an important first step towards strengthening the validity of our programs and in turn, the quality of the surgeons we create.

Bibliography

Accreditation Council for Graduate Medical Education (ACGME) general competencies. (2012)

Available at <http://www.acgme.org>. Accessed November.

Ahmed A, Ishman SL, Laeeq K, Bhatti N. (2013) Assessment of Improvement of Trainee

Surgical Skills in the Operating Room for Tonsillectomy. *Laryngoscope*. 123:1639–1644.

Alderson D. How to critically appraise a research paper. (2012) Developing leadership in

surgical training. *Surgery*. 30:9:479-480.

Anderson CI, Jentz AB, Harkema JM, *et al.* (2005) Assessing the competencies in general

surgery residency training. *American Journal of Surgery*. 189.3:288-292.

doi:10.1016/j.amjsurg.2005.01.001.

Arora S, Sevdalis N, Suliman I, *et al.* (2009) What makes a competent surgeon?: Experts’

and trainees’ perceptions of the roles of a surgeon. *The American Journal of*

Surgery. 198:726–732. doi:10.1016/j.amjsurg.2009.01.015

Banerjee M, Capozzoli M, McSweeney L and Sinha D. (1999) *Beyond kappa. A review of*

inter-rater agreement measures. The Canadian Journal of Statistics. 27.1:3-23.

Baskies MA, Ruchelsman DE, Capeci CM, Zuckerman JD, Egol KA. (2008) Operative

Experience in an Orthopaedic Surgery Residency Program: The Effect of Work-Hour

Restrictions. *The Journal of Bone and Joint Surgery: Topics in Training*. 90:924-7 d

doi:10.2106/JBJS.G.00918.

Benson A, Markwell S, Kohler TS, *et al.* (2012) An operative performance rating system for

urology residents. *Journal of Urology*. 188.5:1877-1882.

doi: 10.1016/j.juro.2012.07.047.

Blum AB, Shea S, Czeisler CA, Landrigan CP, Leape L. (2011) Implementing the 2009 Institute of Medicine recommendations on resident physician work hours, supervision, and safety. *Nature and Science of Sleep*. 3:47–85.

Brieger GH. (1980) Surgery. The Education of American Physicians: Historical Essays. 175-203.

Busari JO, Stammen LA, Gennissen LM, Moonen RM. (2014) Evaluating medical residents as managers of care: a critical appraisal of assessment methods. *Advances in Medical Education and Practice*. 5:27-37.

CanMEDS 2005 Framework. Royal College of Physicians and Surgeons of Canada.

Retrieved from: Accessed on March 4th, 2012.

Canter R. (2011) Impact of reduced working time on surgical training in the United Kingdom and Ireland. *The Surgeon, Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland*. S6-7 doi:10.1016/j.surge.2010.11.020.

Chipman JG, Schmitz CC. (2009) Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. *Journal of the American College of Surgeons* . 209(3):364-370.e2. doi: 10.1016/j.jamcollsurg.2009.05.005.

Chou B, Bowen CW, Handa VL. (2008) Evaluating the competency of gynaecology residents in the operating room: Validation of a new assessment tool. *American Journal of Obstetrics and Gynaecology*. 199(5):571.e1-571.e5. doi: 10.1016/j.ajog.2008.06.082.

- Chou S, Cole G, McLaughlin K, *et al.* (2008) CanMEDS evaluation in Canadian postgraduate training programmes: tools used and programme director satisfaction. *Medical Education*. 42:879–886. doi:10.1111/j.1365-2923.2008.03111.x
- Cicchetti, DV. (1977) Testing the normal approximation and minimal sample size requirements of weighted kappa when number of categories is large. *Applied Psychological Measurement* 5:101-104.
- Cohen J. (1960) A coefficient of agreement for nominal series. *Educational and Psychological Measurement*. 20:37-46.
- Cook DA, Beckman TJ. (2006) Current concepts in validity and reliability for psychometric instruments: theory and application. *American Journal of Medicine*. 119(2):116.e7-16.
- Curran V, Hollet A, Casimiro LM, McCarthy P, Banfield V, and Hall P. (2011) Development and Validation of the Interprofessional Collaborator Assessment Rubric (ICAR). *Interprofessional Care*, 25:339–344
http://www.med.mun.ca/facultypublications/res_view.aspx?ID=138
- Dent JA, Harden RM. (2009) *A Practical Guide for Medical Teacher (3rd ed.)*. Churchill Livingstone: Elsevier.
- Donnon T, Ansari AA, Alawi SA and Violato C. (2014) The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review. *Academic Medicine*. 89:3:511-516.
- Dwyer T, Takahashi SG, Hynes MK, *et al.* (2014) How to assess communication, professionalism, collaboration and the other intrinsic CanMEDs roles in orthopaedic

- residents: use of an objective structured clinical examination (OSCE). *Canadian Journal of Surgery*. 57.4:230-236.
- Feinstein AR, Cicchetti DV. (1990) High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*. 43.6:543-549.
- Ferguson PC, Kraemer W, Nousiainen M, *et al.* (2013) Three-year experience with an Innovative, Modular Competency-Based Curriculum for Orthopaedic Training. *Journal of Bone and Joint Surgery America*. 95:e166:1-6.
- Fitzgibbons SC, Chen J, Jagsi R, Weinstein D. (2012) Long-Term Follow-Up on the Educational Impact of the ACGME Duty Hour Limits. *Annals of Surgery*. 256:1108–1112.
- Frank, JR, Langer B. (2003) “Collaboration, Communication, Management, and Advocacy: Teaching Surgeons New Skills through the CanMEDS Project.” *World Journal of Surgery*. 27:972-8.
- Froelich JM, Milbrant JC, Novicoff WM, *et al.* (2011) Surgical Simulators and Hip Fractures: A Role in Residency Training? *Journal of Surgical Education*. 68.4:298-302.
doi:10.1016/j.jsurg.2011.02.011
- Furey AJ. (2004) The Utility of Classification Systems in Orthopaedic Surgery. Thesis Master of Science. Memorial University of Newfoundland.
- Giesler M, Forster J, Biller S, Fabry G. (2011) Development of a questionnaire to assess medical competencies: Reliability and validity of the questionnaire. *GMS Zeitschrift für Medizinische Ausbildung*. 28.2:1-15.

- Gilbert JH, Orchard C, Bainbridge L, *et al.* (2010) Canadian Interprofessional Health Collaborative: A National Interprofessional Competency Framework.
- Glarner CE, McDonald RJ, Smith AB, *et al.* (2013) Utilizing a Novel Tool for the Comprehensive Assessment of Resident Operative Performance. *Journal of Surgical Education*. 70:6:813-820.
- Goff B, Mandel L, Lentz G, *et al.* (2005) Assessment of resident surgical skills: Is testing feasible? *American Journal of Obstetrics and Gynaecology*. 192.4:1331-8; discussion 1338-40. doi: 10.1016/j.ajog.2004.12.068
- Goff BA, Neilsen PE, Lentz GM, *et al.* (2002) Surgical skills assessment : A blinded examination of obstetrics and gynaecology residents. *American Journal of Obstetrics and Gynaecology*. 186:613-617.
- Golnik KC, Beaver H, Gauba V, *et al.* (2013) Development of a new valid, reliable, and internationally applicable assessment tool of residents' competence in Ophthalmic surgery. *Transactions of the American Ophthalmology Society*. 111:24-33.
- Grantcharov, TP, Reznick RK. (2009) "Training tomorrow's surgeons: what are we looking for and how can we achieve it?" *ANZ Journal of Surgery*. 79:104–107.
- Grober ED, Michael AS, Jewett. (2006) "The concept and trajectory of "operative competence" in surgical training." *Canadian Journal of Surgery*. 49.4.
- Gwet KL. (2012) Handbook of Inter-Rater Reliability. The Definitive Guide to Measuring the Extent of Agreement Among Raters. 3rd edition. Gaithersburg, MD 20886–2696, USA: Advanced Analytics, LLC.

- Hanna WC, Mulder DS, Fried GM, *et al.* (2012) Training Future Surgeons for Management Roles. *Arch Surg.* 147.10:940-944. Published online June 18, 2012.
Doi:10.1001/archsurg.2012.992.
- Hassett JM, Zinnerstrom K, Nawotniak RH, *et al.* (2006) Utilization of standardized patients to evaluate clinical and interpersonal skills of surgical residents. *Surgery.* 140.4:633-639. doi:10.1016/j.surg.2006.07.014
- Hopmans CJ, den Hoed PT, Wallenburg I, *et al.* (2013) Surgeons' Attitude Toward a Competency-Based Training and Assessment Program: Results of a Multicenter Survey. *Journal of Surgical Education.* 70.5:647-654.
- Hsieh FY, Bloch DA, Larsen MD. (1998) A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine.* 17.14:1623-34.
- Hulley SB, Cummings SR. (1988) "Designing clinical research: An epidemiological approach." Williams and Wilkins, Baltimore. 31-42.
- Humphrey-Murto S, Khalidi N, Smith D, *et al.* (2009) Resident Evaluations: The Use of Daily Evaluation Forms in Rheumatology Ambulatory Care. *The Journal of Rheumatology.* 36.6:1298-1303.
- Ishman SL, Benke JR, Johnson KE, *et al.* (2012) Blinded evaluation of inter-rater reliability of an operative competency assessment tool for direct laryngoscopy and rigid bronchoscopy. *Archives of Otolaryngology Head and Neck Surgery.* 1-7. doi: 10.1001/2013.jamaoto.115.

- Ishman SL, Brown DJ, Boss EF. (2010) Development and Pilot Testing of an Operative Competency Assessment Tool for Paediatric Direct Laryngoscopy and Rigid Bronchoscopy. *Laryngoscope*. 11.120:2294-2300. doi: 10.1002/lary.21067.
- Jefferies AB, Simmons D, Tabak, *et al.* (2007) Using an objective structured clinical examination (OSCE) to assess multiple physician competencies in postgraduate training. *Medical Teacher*. 2.3:183-91.
- Johnson D, Cujec B. (1998) Comparison of Self, Nurse, and Physician Assessment of Residents Rotating Through an Intensive Care Unit. *Critical Care Medicine*. 26.11:1811-6.
- Kamath AF, Bladwin K, Meade LK, Powell AC, Mehta S. (2011) The Increased Burden of Further Proposed Orthopaedic Resident Work-Hour Restrictions. *The Journal of Bone and Joint Surgery: Topics in Training*. 93.31:1-8. doi:10.2106/JBJS.I.01676.
- Kassam A, Donnon T, Rigby I. (2014) Validity and reliability of an in-training evaluation report to measure the CanMEDs roles in emergency medicine resident. *Canadian Journal of Emergency Medicine*. 16.2:144-150.
- Koehler RJ, Nicandri GT. (2013) Using the Arthroscopic Surgery Skill Evaluation Tool as a Pass-Fail Examination. *Journal of Bone and Joint Surgery America*. 95.187:1-6.
<http://dx.doi.org/10.2106/JBJS.M.00340>
- Kramer MS, Feinstein AR. (1981) The Biostatistics of Concordance. *Clinical Pharmacology and Therapeutics*. 29:111-123.

- Landis RJ, Koch GG. (1977) The measurement of observer agreement for categorical data. *Biometrics*. 33:159-174.
- Laeq K., Bhatti NI, Carey JP, *et al.* (2009) Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope*. 119.12:2402-2410. doi: 10.1002/lary.20678.
- Laeq K, Infusino S, Lin SY, *et al.* (2010) Video-based assessment of operative competency in endoscopic sinus surgery. *American Journal of Rhinology and Allergy*. 24.3:234-237. doi: 10.2500/ajra.2010.24.3434.
- Larson JL, Williams RG, Ketchum J, *et al.* (2005) Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. 138.4:640-7; discussion 647-9. doi: 10.1016/j.surg.2005.07.017.
- Lentz GM, Mandel LS, Lee D, *et al.* (2001) Testing surgical skills of obstetric and gynaecologic residents in a bench laboratory setting: Validity and reliability. *American Journal of Obstetrics and Gynaecology*. 184.7:1462-8; discussion 1468-70.
- Lin SY, Laeq K, Ishii M, *et al.* (2009) Development and pilot-testing of a feasible, reliable, and valid operative competency assessment tool for endoscopic sinus surgery. *American Journal of Rhinology and Allergy*. 23.3:354-359. doi:10.2500/ajra.2009.23.3275.
- Martin JA, Regehr G, Reznick R, *et al.* (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*. 84.2:273-278.
- Massagli TL, Carline JD. (2007) Reliability of a 360-Degree Evaluation to Assess Resident Competence. *American Journal of Physical Medicine and Rehabilitation*. 86:845-852.

- McGinn T, Wyer PC; Newman TB; *et al.* (2004) Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *Canadian Medical Association Journal*. 171:11.
- Mickelson JJ, MacNeily AE. (2008) Translational education: tools for implementing the CanMEDS competencies in Canadian urology residency training. *Canadian Urological Association Journal*. 2:4.
- Moktar J, Popkin CA, Howard A, Murnaghan, ML. (2014) Development of a Cast Application Simulator and Evaluation of Objective Measures of Performance. *Journal of Bone and Joint Surgery America*. 96:76:1-6.
- Nauta RJ. (2012) Residency Training Oversight(s) in Surgery: The History and Legacy of the Accreditation Council for Graduate Medical Education Reforms. *Surgical Clinics of North America*. 92:117–123.
- Neira VM, Bould MD, Nakajima A, *et al.* (2013) “GIOSAT”: a tool to assess competencies during simulated crisis. *Canadian Journal of Anaesthesia*. 60:280-289.
- Ortwein H, Knigge M, Rehberg B, *et al.* (2011) “Validation of core competencies during residency training in anaesthesiology.” *German Medical Science*, 9.
- Palter VN, Grantcharov TP. (2012) A prospective study demonstrating the reliability and validity of two procedure-specific evaluation tools to assess operative competency in laparoscopic colorectal surgery. *Surgical Endoscopy*. 26:2489-2503.

- Parent F, Jouquan J, de Kettle JM. (2013) CanMEDs and other “competency and outcome-based approaches” in medical education: clarifying the ongoing ambiguity. *Advances in Health Science Education*. 18:115–122. doi:10.1007/s10459-012-9402-z.
- Reznick, R, Regehr G., MacRae H. (1997) Testing Technical Skills via and Innovative “Bench Station” Examination. *The American Journal of Surgery*. 173:226-230.
- Roberson DW, Kentala E, Forbes P. (2005) Development and Validation of an Objective Instrument to Measure Surgical Performance in Tonsillectomy. *Laryngoscope*. 115:2127-2137. doi: 10.1097/01.mlg.0000178329.23359.30.
- Rodriguez E, Seigelman J, Leone K, Kessler C. (2012) Assessing Professionalism: Summary of the Working Group on Assessment of Observable Learner Performance. *Academic Emergency Medicine*. 19:1372-1378.
- Rose SH, Long TR, Elliott BA, Brown MJ. (2009) A Historical Perspective on Resident Evaluation, the Accreditation Council for Graduate Medical Education Outcome Project and Accreditation for Graduate Medical Education Duty Hour Requirement. *International Anaesthesia Research Society*. 109.1:190-193.
- Sackett DL, Haynes DB, Guyatt GH, Tugwell P. (1991) Clinical Epidemiology: A basic science for clinical medicine. Boston etc: Lippincott-Raven. 51-153.
- Schmitt N. (1996) Uses and abuses of coefficient alpha. *Psychological Assessment*. 8:350-3.
- Sherbino J, Kulasegarm K, Worster A, Norman GR. (2013) The reliability of encounter cards to assess the CanMEDs roles. *Advances in Health Science Education*. 18:987-996.

Sijtsma, K. (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha.

Psychometrika. 74.1:107-120. doi: 10.1007/S11336-008-9101-0

Stack BC, Seigel E, Bodenner D, *et al.* (2010) A study of resident proficiency with thyroid

surgery: creation of a thyroid specific tool. *Otolaryngology and Head and Neck Surgery*

142.6:856-62.

Tavakol M, Dennick R. (2011) Making sense of Cronbach's alpha. *International Journal of*

Medical Education. 2:53-55 DOI: 10.5116/ijme.4dfb.8dfd.

Thorndike, EL. (1920) "A constant error in psychological ratings." *Journal of applied*

psychology. 4.1:25-29.

Tomisato S, Venter J, Weller J, Drachman D. (2013) Evaluating the Utility, Reliability, and

Validity of a Resident Performance Evaluation Instrument. *Academic Psychiatry*. 38:458-

463.

Trajkovski T, Veillette C, Backstein D, Wadey VM, Kraemer B. (2012) Resident self

assessment of operative experience in primary total knee and total hip arthroplasty: Is it

accurate? *Canadian Journal of Surgery*. 55.4.2:153-157.

Wanzel KR, Ward M, Reznick RK. (2002) Teaching the surgical craft: from selection to

certification. *Current Problems in Surgery*. 39:574-659.

Williams RG, Sanfey H, Chen X, Dunnington GL. (2012) A Controlled Study to Determine

Measurement Conditions Necessary for a Reliable and Valid Operative Performance

Assessment. *Annals of Surgery*. 256:1.177-187.

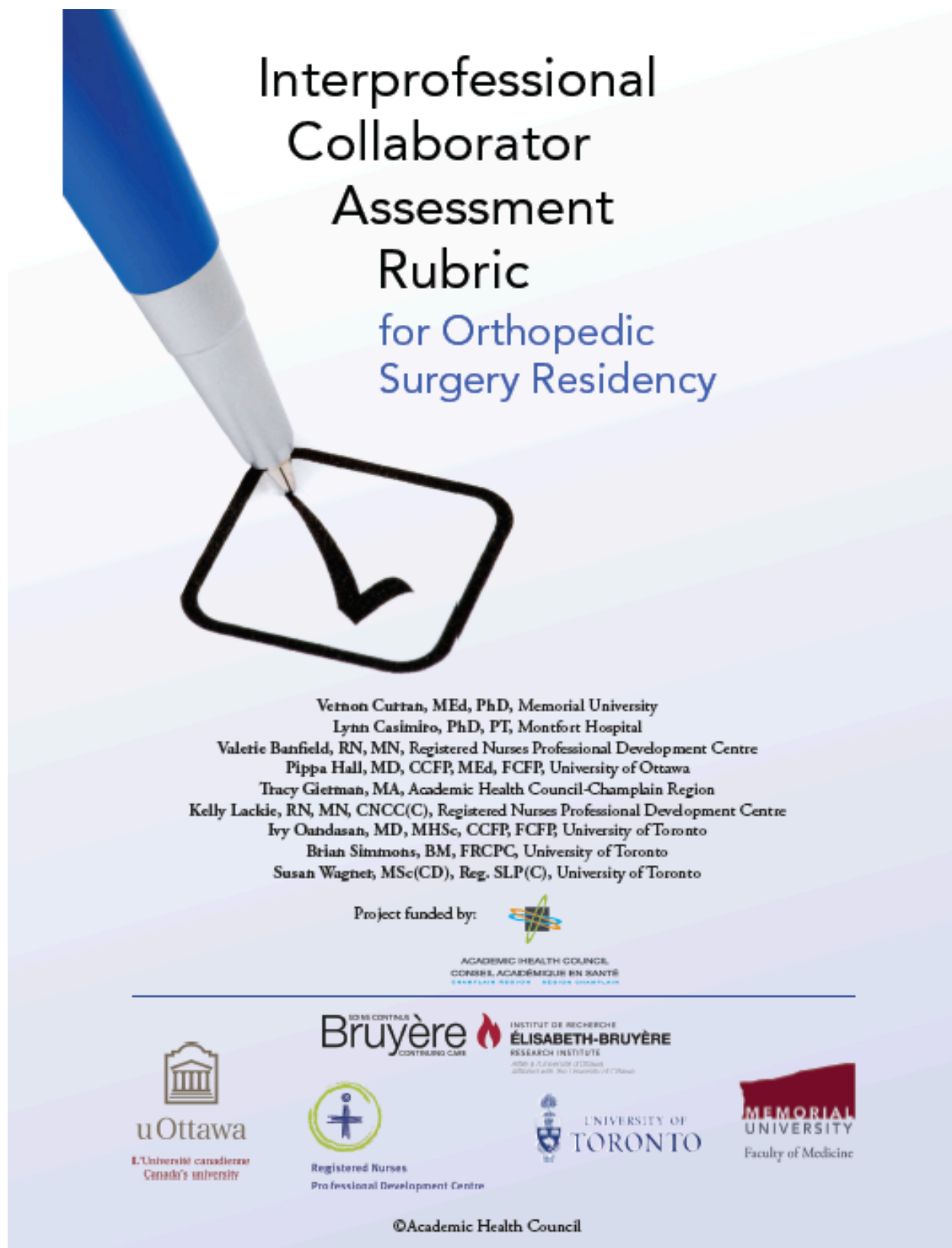
Winckel CP, Reznick RK, Cohen R. (2004) Reliability and Construct Validity of a Structured Technical Skills Assessment Form. *The American Journal of Surgery*. 167:423-427.

Appendix A: Summary of surgical evaluation studies

	Type of Study (Prospective vs cross-sectional)	n = (eval)	Outcome by Statistical Methods (Reliability and Validity measures)
Cast Application Moktar et al 2014	Cross-sectional cohort	9	Intraclass correlation
Tonsillectomy trainee skills Ahmed 2013	Prospective longitudinal validation	83	Comparison of mean scores Cronbach's alpha
Ophthalmic Eval Golnik et al 2013	Cross-sectional cohort	6	Cronbach's alpha
Arthroscopic Skills Eval Koehler et al 2013	Cross-sectional cohort	60	Intraclass correlation
Resident Operative Performance Glärner et al 2013	Cross-sectional cohort	63	ANOVA
Manager Role Hanna et al 2012	Course evaluation	43	Self evaluation
An operative performance rating system Benson et al. 2012	Cross-sectional cohort	175	Cronbach's alpha ANOVA
A prospective study Palter et al. 2012	Cross-sectional cohort	43	Cronbach's alpha Construct validity – Mann Whitney U test
Video based assessment of operative competency Laeq et al. 2012	Cross-sectional cohort, blinded	40	Cronbach's alpha Interclass correlation coefficients Construct validity - ANOVA
Blinded evaluation of inter-rater reliability Ishman et al. 2012	Cross-sectional cohort, blinded	45	Kappa Intraclass correlation Cronbach's alpha
A study of resident proficiency Stack et al. 2010	Cross-sectional cohort	97	Cronbach's alpha Construct validity ANOVA
Development and Pilot Testing Ishman et al. 2010	Cross-sectional cohort	44	Cronbach's alpha Kappa
Using objective structured assessment Chipman et al. 2009	Cross-sectional cohort	38	Cronbach's alpha ANOVA
Development and Pilot testing sinus surgery Lin et al. 2009	Cross-sectional cohort	51	Construct validity – by percentage scores Cronbach's alpha


Pilot testing of an assessment Laeq et al. 2009	Cross-sectional cohort, blinded	118	Percent agreement Construct validity – ANOVA
Evaluating the competency of gynecology Chou et al. 2008	Cross-sectional cohort	362	Construct validity – factor analysis Cronbach's alpha Pearson's item-total correlation ANOVA
Development and Validation Roberson et al. 2005	Cross-sectional cohort	55	Weighted kappa Face validity Construct validity – absolute scores
Feasibility, reliability and validity Larson et al. 2005	Cross-sectional cohort, blinding	77	Construct validity - ANOVA Absolute scores
Assessment of resident surgical skills Goff et al. 2004	Cross-sectional cohort, blinded, multi-center	116	Cronbach's alpha Intraclass correlation coefficients Construct validity – ANOVA
Surgical Skills Assessment Goff et al. 2002	Cross-sectional cohort	102	Cronbach's alpha Intraclass correlation coefficients
Testing surgical skills of obstetric Lentz et al. 2001	Cross-sectional cohort	180	Cronbach's alpha Intraclass correlation coefficients Construct validity – ANOVA
Objective structured assessment Martin et al. 1997	Cross-sectional cohort	20	ANOVA Cronbach's alpha Intraclass correlation coefficient Construct validity MANOVA
Testing technical skill Reznick et al. 1996	Cross-sectional cohort	384	ANOVA Cronbach's alpha

Appendix B: Interprofessional Collaborator Assessment Rubric – Orthopaedic surgery








Interprofessional Collaborator Assessment Rubric for Orthopaedic Surgery Residency

Vernon Cuttari, MEd, PhD, Memorial University
Lynn Casimiro, PhD, PT, Montfort Hospital
Valette Banfield, RN, MN, Registered Nurses Professional Development Centre
Pippa Hall, MD, CCFP, MEd, FCFP, University of Ottawa
Tracy Gierman, MA, Academic Health Council-Champlain Region
Kelly Lackie, RN, MN, CNCC(C), Registered Nurses Professional Development Centre
Ivy Oubdasah, MD, MHSc, CCFP, FCFP, University of Toronto
Brian Simmons, BM, FRCPC, University of Toronto
Susan Wagnet, MSc(CD), Reg. SLP(C), University of Toronto

Project funded by: 

ACADEMIC HEALTH COUNCIL
CONSEIL ACADÉMIQUE EN SANTÉ
CHAMPLAIN REGION - RÉGION CHAMPLAIN



©Academic Health Council

What is a Rubric?

A Rubric is an assessment tool that lists a set of performance criteria which define and describe the important competencies being assessed. Rubrics are useful to instructors because it can improve the planning of learning experiences and increase the quality of direct instruction by providing focus, emphasis, and attention to particular details as a model for learners.

For learners, a rubric provides clear targets of proficiency to aim for. Learners can use Rubrics for self-assessment as

individuals, in groups, and for peer assessment. It is believed that Rubrics may improve learners' performance and therefore increase learning, particularly when learners receive Rubrics beforehand, understand how they will be evaluated and can prepare accordingly. Rubrics are becoming increasingly popular with educators moving toward more authentic, performance-based assessments.

Interprofessional Collaborator Assessment Rubric

Competency Category	Descriptor																									
Conflict Management/Resolution:	Ability to effectively manage and resolve conflict between and with other providers, patients/clients and families.																									
Competency Statements	<div>1. Demonstrates active listening and is respectful of different perspectives and opinions from others</div> <div>2. Works with others to manage and resolve conflict effectively.</div>																									
Rubric Scale																										
	<table><tr><th>Not Observable</th><th>Minimal</th><th>Developing</th><th>Competent</th><th>Mastery</th></tr><tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td></tr><tr><td>Dimensions</td><td></td><td></td><td></td><td></td></tr><tr><td>Active Listening</td><td><div><input type="checkbox"/> Does not use active listening techniques when others are speaking.</div></td><td><div><input type="checkbox"/> Occasionally uses active listening when others are speaking.</div></td><td><div><input type="checkbox"/> Frequently uses active listening when others are speaking.</div></td><td><div><input type="checkbox"/> Consistently uses active listening when others are speaking.</div></td></tr><tr><td>Behavioural Indicator</td><td></td><td></td><td></td><td></td></tr></table>	Not Observable	Minimal	Developing	Competent	Mastery		1	2	3	4	Dimensions					Active Listening	<div><input type="checkbox"/> Does not use active listening techniques when others are speaking.</div>	<div><input type="checkbox"/> Occasionally uses active listening when others are speaking.</div>	<div><input type="checkbox"/> Frequently uses active listening when others are speaking.</div>	<div><input type="checkbox"/> Consistently uses active listening when others are speaking.</div>	Behavioural Indicator				
Not Observable	Minimal	Developing	Competent	Mastery																						
	1	2	3	4																						
Dimensions																										
Active Listening	<div><input type="checkbox"/> Does not use active listening techniques when others are speaking.</div>	<div><input type="checkbox"/> Occasionally uses active listening when others are speaking.</div>	<div><input type="checkbox"/> Frequently uses active listening when others are speaking.</div>	<div><input type="checkbox"/> Consistently uses active listening when others are speaking.</div>																						
Behavioural Indicator																										

Using the Collaborator Rubric

The Interprofessional Collaborator Assessment Rubric is intended for use in the assessment of interprofessional collaborator competencies. Collaborative practice in health care occurs when multiple health workers from different professional backgrounds provide comprehensive services by working with patients, their families, carers and communities to deliver the highest quality of care across settings (WHO, 2010)¹. Development of the Rubric tool was guided by an interprofessional advisory committee comprising educators from the fields of medicine, nursing and the rehabilitative sciences.

Key Principles

- 1) The Rubric has been developed for usage across different health professional education programs and in different learning contexts.
- 2) The Rubric dimensions are not intended to coincide with a specific year or level of a learner in his/her program of studies.
- 3) The Rubric may be used as a tool for formative and summative assessment of learners' competencies in

interprofessional collaboration. As a formative assessment, the Rubric would allow learners to receive constructive feedback on competency areas for further development and improvement. As a summative assessment, the Rubric may be used to assess learners' achievement. The Rubric may also be introduced early in a program and used repeatedly to assess growth and development over time.

- 4) Usage of the Rubric in a reliable manner may require multiple interactions and repeated observation of a learner over a period of time.
- 5) Programs/disciplines should define remediation opportunities for learners not achieving an acceptable level of competency within their program area.

Rubric Validity

The Rubric dimensions are based on interprofessional collaborator competency statements that were developed and validated through a typological analysis of national and international competency frameworks, a Delphi survey of experts, and interprofessional focus groups with students and faculty.

¹ World Health Organization (WHO) Study Group on Interprofessional Education and Collaborative Practice. (2010). *Framework for Action on Interprofessional Education & Collaborative Practice*. Geneva, Switzerland: World Health Organization.

Interprofessional Collaborator Assessment Rubric

Instructions: For each of the dimensions below, check specific phrases which describe the performance of the learner.

Notes

Assess by what is appropriate to the context/task.

- Occasionally: the learner demonstrates the desired behaviour once in a while.
- Frequently: the learner demonstrates the desired behaviour most of the time.
- Consistently: the learner always demonstrates the desired behaviour.

Communication: Ability to communicate effectively in a respectful and responsive manner with others ("others" includes team members, patient/client, and health providers outside the team).

1. Communicates and expresses ideas in an assertive and respectful manner.
2. Uses communication strategies (e.g. oral, written, information technology) in an effective manner with others.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
<i>Respectful Communication</i>		<input type="checkbox"/> Communicates with others in a disrespectful manner.	<input type="checkbox"/> Occasionally communicates with others in a confident, assertive and respectful manner.	<input type="checkbox"/> Frequently communicates with others in a confident, assertive and respectful manner.	<input type="checkbox"/> Consistently communicates with others in a confident, assertive and respectful manner.
		<input type="checkbox"/> Does not communicate opinion or pertinent views on patient care with others.	<input type="checkbox"/> Occasionally communicates opinion or pertinent views on patient care with others.	<input type="checkbox"/> Frequently communicates opinion and pertinent views on patient care with others.	<input type="checkbox"/> Consistently communicates opinion and pertinent views on patient care with others.
		<input type="checkbox"/> Does not respond or reply to requests.	<input type="checkbox"/> Occasionally responds or replies to requests in a timely manner.	<input type="checkbox"/> Frequently responds or replies to requests in a timely manner.	<input type="checkbox"/> Consistently responds or replies to requests in a timely manner.
<i>Communication Strategies</i>		<input type="checkbox"/> Communication is illogical and unstructured.	<input type="checkbox"/> Occasionally communicates in a logical and structured manner.	<input type="checkbox"/> Frequently communicates in a logical and structured manner.	<input type="checkbox"/> Consistently communicates in a logical and structured manner.
		<input type="checkbox"/> Does not explain discipline-specific terminology/jargon.	<input type="checkbox"/> Occasionally explains discipline-specific terminology/jargon.	<input type="checkbox"/> Frequently explains discipline-specific terminology/jargon.	<input type="checkbox"/> Consistently explains discipline-specific terminology/jargon.
		<input type="checkbox"/> Does not use strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).	<input type="checkbox"/> Occasionally uses strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).	<input type="checkbox"/> Frequently uses strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).	<input type="checkbox"/> Consistently uses strategies that are appropriate for communicating with individuals with impairments (e.g., hearing, cognitive).
Comments:					

Collaboration: Ability to establish/maintain collaborative working relationships with other providers, patients/clients and families.

1. Establishes collaborative relationships with others in planning and providing patient/client care.
2. Promotes the integration of information from others in planning and providing care for patients/clients.
3. Upon approval of the patient/client or designated decision-maker, ensures that appropriate information is shared with other providers.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
<i>Collaborative Relationship</i>		<input type="checkbox"/> Does not establish collaborative relationships with others.	<input type="checkbox"/> Occasionally establishes collaborative relationships with others.	<input type="checkbox"/> Frequently establishes collaborative relationships with others.	<input type="checkbox"/> Consistently establishes collaborative relationships with others.
<i>Integration of Information from others</i>		<input type="checkbox"/> Does not integrate information from others in planning and providing patient/client care.	<input type="checkbox"/> Occasionally integrates information from others in planning and providing patient/client care.	<input type="checkbox"/> Frequently integrates information and perspectives from others in planning and providing patient/client care.	<input type="checkbox"/> Consistently integrates information and perspectives from others in planning and providing patient/client care.
<i>Information Sharing</i>		<input type="checkbox"/> Does not share information with other providers.	<input type="checkbox"/> Occasionally shares information with other providers that is useful for the delivery of patient/client care.	<input type="checkbox"/> Frequently shares information with other providers that is useful for the delivery of patient/client care.	<input type="checkbox"/> Consistently shares information with other providers that is useful for the delivery of patient/client care.
		<input type="checkbox"/> Does not seek approval of patient/client or designated decision-maker when information is shared.	<input type="checkbox"/> Occasionally seeks approval of the patient/client or designated decision-maker when information is shared.	<input type="checkbox"/> Frequently seeks approval of the patient/client or designated decision-maker when information is shared.	<input type="checkbox"/> Consistently seeks approval of the patient/client or designated decision-maker when information is shared.
Comments:					

Roles and Responsibility: Ability to explain one's own roles and responsibilities related to patient/client and family care (e.g. scope of practice, legal and ethical responsibilities); and to demonstrate an understanding of the roles, responsibilities and relationships of others within the team.

1. Describes one's own roles and responsibilities in a clear manner.
2. Integrates the roles and responsibilities of others with one's own to optimize patient/client care.
3. Accepts accountability for one's contributions.
4. Shares evidence-based and/or best practice discipline-specific knowledge with others.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
<i>Roles and Responsibilities</i>		<input type="checkbox"/> Does not describe one's own role and responsibilities with the team/patient/family.	<input type="checkbox"/> Occasionally describes one's own role and responsibilities with the team/patient/family.	<input type="checkbox"/> Frequently describes one's own roles and responsibilities with the team/patient/family.	<input type="checkbox"/> Consistently describes one's own roles and responsibilities in a clear manner with the team/patient/family.
<i>Role/Responsibility Integration</i>		<input type="checkbox"/> Does not include the roles and responsibilities of other providers in the delivery of patient care.	<input type="checkbox"/> Occasionally includes the roles and responsibilities of other providers in the delivery of patient care.	<input type="checkbox"/> Frequently includes the roles and responsibilities of all necessary health providers to optimize collaborative patient/client care.	<input type="checkbox"/> Consistently promotes and includes the roles and responsibilities of all necessary health providers to optimize collaborative patient/client care.
<i>Accountability</i>		<input type="checkbox"/> Does not demonstrate professional judgment when assuming tasks or delegating tasks.	<input type="checkbox"/> Occasionally demonstrates professional judgment when assuming tasks or delegating tasks.	<input type="checkbox"/> Frequently demonstrates professional judgment when assuming tasks or delegating tasks.	<input type="checkbox"/> Consistently demonstrates professional judgment when assuming tasks or delegating tasks.
		<input type="checkbox"/> Does not accept responsibility for individual actions that impact the team.	<input type="checkbox"/> Occasionally accepts responsibility for individual actions that impact the team.	<input type="checkbox"/> Frequently accepts responsibility for individual actions that impact the team.	<input type="checkbox"/> Consistently accepts responsibility for individual actions that impact the team.
<i>Sharing Evidence-Based/ Best Practice Knowledge</i>		<input type="checkbox"/> Does not share evidence-based or best practice discipline-specific knowledge with others.	<input type="checkbox"/> Occasionally shares evidence-based or best practice discipline-specific knowledge with others.	<input type="checkbox"/> Frequently shares evidence-based or best practice discipline-specific knowledge with others.	<input type="checkbox"/> Consistently shares evidence-based or best practice discipline-specific knowledge with others.
Comments:					

Collaborative Patient/Client-Family Centred Approach: Ability to apply patient/client-centred principles through interprofessional collaboration.

1. Seeks input from patient/client and family in a respectful manner regarding feelings, beliefs, needs and care goals.
2. Integrates patient's/client's and family's life circumstances, cultural preferences, values, expressed needs, and health beliefs/behaviours into care plans.
3. Shares options and health care information with patients/clients and families.
4. Advocates for patient/client and family as partners in decision-making processes.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
<i>Patient/Client Input</i>		<input type="checkbox"/> Does not seek input from patient/client and family.	<input type="checkbox"/> Occasionally seeks input from patient/client and family.	<input type="checkbox"/> Frequently seeks input from patient/client and family.	<input type="checkbox"/> Consistently seeks input from patient/client and family.
<i>Integration of Patient/Client Beliefs and Values</i>		<input type="checkbox"/> Does not integrate patient's/client's and family's circumstances, beliefs and values into care plans.	<input type="checkbox"/> Occasionally integrates the patient's/client's and family's circumstances, beliefs and values into care plans.	<input type="checkbox"/> Frequently integrates patient's/client's and family's circumstances, beliefs and values into care plans.	<input type="checkbox"/> Consistently promotes and integrates patient's/client's and family's circumstances, beliefs and values into care plans.
<i>Information Sharing with Patient/Client</i>		<input type="checkbox"/> Does not share options and health care information with patients/clients and families.	<input type="checkbox"/> Occasionally shares options and health care information with patients/clients and families.	<input type="checkbox"/> Frequently shares options and health care information with patients/clients and families.	<input type="checkbox"/> Consistently shares options and health care information with patients/clients and families.
<i>Patient Advocacy in Decision-Making</i>		<input type="checkbox"/> Does not advocate for patient/client and family as partners in decision-making processes.	<input type="checkbox"/> Occasionally advocates for patient/client and family as partners in decision-making processes.	<input type="checkbox"/> Frequently advocates for patient/client and family as partners in decision-making processes.	<input type="checkbox"/> Consistently advocates for patient/client and family as partners in decision-making processes.
Comments:					

Team Functioning: Ability to contribute to effective team functioning to improve collaboration and quality of care.

1. Recognizes and contributes to effective team functioning and dynamics.
2. Recognizes that leadership within the healthcare team may alternate or be shared depending on the situation.
3. Contributes in interprofessional team discussions.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
<i>Team Functioning and Dynamics</i>		<input type="checkbox"/> Does not recognize strategies that will improve team functioning.	<input type="checkbox"/> Occasionally demonstrates recognition of strategies that will improve team functioning.	<input type="checkbox"/> Frequently demonstrates recognition of strategies that will improve team functioning.	<input type="checkbox"/> Consistently demonstrates recognition of strategies that will improve team functioning.
<i>Shared Leadership</i>		<input type="checkbox"/> Does not recognize the importance of alternating or sharing leadership with others.	<input type="checkbox"/> Occasionally shares leadership and alternates leadership with others when appropriate for the discipline involved.	<input type="checkbox"/> Frequently shares leadership and alternates leadership with others when appropriate for the discipline involved.	<input type="checkbox"/> Consistently shares leadership and alternates leadership with others when appropriate for the discipline involved.
<i>Team Discussion</i>		<input type="checkbox"/> Does not contribute to interprofessional team discussions.	<input type="checkbox"/> Occasionally contributes to interprofessional team discussions.	<input type="checkbox"/> Frequently contributes to interprofessional team discussions.	<input type="checkbox"/> Consistently contributes to interprofessional team discussions.
Comments:					

Conflict Management/Resolution: Ability to effectively manage and resolve conflict between and with other providers, patients/clients and families.

1. Demonstrates active listening and is respectful of different perspectives and opinions from others.
2. Works with others to manage and resolve conflict effectively.

Dimensions	Not Observable	Minimal 1	Developing 2	Competent 3	Mastery 4
<i>Respect for different perspectives</i>		<input type="checkbox"/> Does not consider the perspectives and opinions of others.	<input type="checkbox"/> Occasionally seeks the perspectives and opinions of others.	<input type="checkbox"/> Frequently seeks the perspectives and opinions of others.	<input type="checkbox"/> Consistently seeks the perspectives and opinions of others.
		<input type="checkbox"/> Does not seek clarification in a respectful manner when misunderstandings arise.	<input type="checkbox"/> Occasionally seeks clarification when misunderstandings arise, but it is not necessarily done in a respectful manner.	<input type="checkbox"/> Frequently seeks clarification in a respectful manner when misunderstandings arise.	<input type="checkbox"/> Consistently seeks clarification in a respectful manner when misunderstandings arise.
<i>Conflict Management</i>		<input type="checkbox"/> Does not manage or resolve conflict with others.	<input type="checkbox"/> Occasionally uses appropriate conflict resolution strategies to manage and/or resolve conflict.	<input type="checkbox"/> Frequently uses appropriate conflict resolution strategies to manage and/or resolve conflict.	<input type="checkbox"/> Consistently uses appropriate conflict resolution strategies to manage and/or resolve conflict.
Comments:					

Appendix C: Health Research Ethics Board Approval for ICAR



**Ethics Office
Suite 200, Eastern Trust Building
95 Bonaventure Avenue
St. John's, NL
A1B 2X5**

March 27, 2012

Dr. Nicholas Smith
Orthopedic Surgery
Memorial University

Dear Dr. Smith

Reference # 12.059

**Re: Evaluating Reliability of the Interprofessional Collaborator Assessment Rubric (ICAR)
for the CanMEDS Collaborator Role in an Orthopedics Residency Program**

Your application received an expedited review by a Sub-Committee of the Health Research Ethics Board and **full approval** was granted effective **March 27, 2012**.

This approval will lapse on **March 26, 2013**. It is your responsibility to ensure that the Ethics Renewal form is forwarded to the HREB office prior to the renewal date. *The information provided in this form must be current to the time of submission and submitted to the HREB not less than 30 nor more than 45 days of the anniversary of your approval date.* The Ethics Renewal form can be downloaded from the HREB website <http://www.hrea.ca>.

This is to confirm that the following documents have been reviewed and approved or acknowledged (as indicated):

- Application, approved
- Proposal, approved
- Revised Informed Consent Form, approved

The Health Research Ethics Board advises THAT IF YOU DO NOT return the completed Ethics Renewal form prior to date of renewal:

- *Your ethics approval will lapse*
- *You will be required to stop research activity immediately*
- *You may not be permitted to restart the study until you reapply for and receive approval to undertake the study again*

Lapse in ethics approval may result in interruption or termination of funding

email: info@hrea.ca

Phone: 777-8949

FAX: 777-8776

/

It is your responsibility to seek the necessary approval from the Regional Health Authority or other organization as appropriate.

Modifications of the protocol/consent are not permitted without prior approval from the Health Research Ethics Board. Implementing changes in the protocol/consent without HREB approval may result in the approval of your research study being revoked, necessitating cessation of all related research activity. Request for modification to the protocol/consent must be outlined on an amendment form (available on the HREB website) and submitted to the HREB for review.

This research ethics board (the HREB) has reviewed and approved the research protocol and documentation as noted above for the study which is to be conducted by you as the qualified investigator named above at the specified site. This approval and the views of this Research Ethics Board have been documented in writing. In addition, please be advised that the Health Research Ethics Board currently operates according to *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*; *ICH Guidance E6: Good Clinical Practice* and applicable laws and regulations. The membership of this research ethics board is constituted in compliance with the membership requirements for research ethics boards as defined by *Health Canada Food and Drug Regulations Division 5; Part C*

Notwithstanding the approval of the HREB, the primary responsibility for the ethical conduct of the investigation remains with you.

We wish you every success with your study.

Sincerely,



Patricia Grainger, Acting Chair
Non-Clinical Trials
Health Research Ethics Board

C VP Research c/o Office of Research, MUN
VP Research c/o Patient Research Centre, Eastern Health
HREB meeting date: April 5, 2012

Appendix D: Surgical Encounters Form

Surgical Encounter Form

Staff Name:

Resident Name:

Date:

	DOES NOT MEET	MEETS	EXCEEDS	NOT OBSERVED
Medical Expert				
The resident showed competency in the following:				
Indications for surgery	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Understanding the risks involved with surgical procedure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Understands the contraindications for surgery	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Technical Skills				
Resident ordered the appropriate equipment pre-operatively	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Resident displayed an understanding of the various equipment/hardware that would be used	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Competent in the surgical approach and dissection required	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Competent in the sequential steps of the procedure and understood the order in which they were performed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

	DOES NOT MEET	MEETS	EXCEEDS	NOT OBSERVED
Resident dealt with any complications of the procedure that arose	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Resident displayed competence in this procedure technically and would be safe to perform it	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Communicator				
Resident spoke with the patient pre-operatively explaining risks and benefits of the operation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Resident was competent in answering questions posed by the patient or family members	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Collaborator				
Resident worked with the other health care provider providing any pre-operative concerns to the anesthetic team members	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Any medical consultation that was required pre-operatively was performed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Advocate				
Resident advocated for timely access to OR recognizing urgent versus elective timing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Resident advocated for any pre-operative consultations to be performed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Comments:

Appendix E: Health Research Ethics Board Approval for SEF



**Ethics Office
Suite 200, Eastern Trust Building
95 Bonaventure Avenue
St. John's, NL
A1B 2X5**

January 23, 2013

Dr. Nicholas Smith
9 Turnberry Street
St. John's, NL A1A 5P3

Dear Dr Smith

Reference #13.011

Re: Evaluating the Reliability and Validity of an Orthopedic Surgical Assessment Tool

Your application received an expedited review by a Sub-Committee of the Health Research Ethics Board and **full approval** was granted effective **January 22, 2013**.

This approval will lapse on **January 21, 2014**. It is your responsibility to ensure that the Ethics Renewal form is forwarded to the HREB office prior to the renewal date. *The information provided in this form must be current to the time of submission and submitted to the HREB not less than 30 nor more than 45 days of the anniversary of your approval date.* The Ethics Renewal form can be downloaded from the HREB website <http://www.hrea.ca>.

This is to confirm that the following documents have been reviewed and approved or acknowledged (as indicated):

- Application, approved
- Revised consent form

The Health Research Ethics Board advises THAT IF YOU DO NOT return the completed Ethics Renewal form prior to date of renewal:

- *Your ethics approval will lapse*
- *You will be required to stop research activity immediately*
- *You may not be permitted to restart the study until you reapply for and receive approval to undertake the study again*

Lapse in ethics approval may result in interruption or termination of funding

It is **your responsibility to seek the necessary approval from the Regional Health Authority or other organization as appropriate.**

email: info@hrea.ca

Phone: 777-8949

FAX: 777-8776

Modifications of the protocol/consent are not permitted without prior approval from the Health Research Ethics Board. Implementing changes in the protocol/consent without HREB approval may result in the approval of your research study being revoked, necessitating cessation of all related research activity. Request for modification to the protocol/consent must be outlined on an amendment form (available on the HREB website) and submitted to the HREB for review.

This research ethics board (the HREB) has reviewed and approved the research protocol and documentation as noted above for the study which is to be conducted by you as the qualified investigator named above at the specified site. This approval and the views of this Research Ethics Board have been documented in writing. In addition, please be advised that the Health Research Ethics Board currently operates according to *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans; ICH Guidance E6: Good Clinical Practice* and applicable laws and regulations. The membership of this research ethics board is constituted in compliance with the membership requirements for research ethics boards as defined by *Health Canada Food and Drug Regulations Division 5; Part C*. Notwithstanding the approval of the HREB, the primary responsibility for the ethical conduct of the investigation remains with you.

We wish you every success with your study.

Sincerely,



Dr. Fern Brunger
Chair
Non-Clinical Trials
Health Research Ethics Board

C VP Research c/o Office of Research, MUN
VP Research c/o Patient Research Centre, Eastern Health
HREB meeting date: February 7, 2013

Appendix F: Collaborator Role Abstract

Introduction: To determine the reliability of an assessment method of the Collaborator role within an orthopaedic surgery residency program as defined by the Royal College of Physicians and Surgeons of Canada, CanMEDs framework. **Methods:** A critical appraisal was undertaken that indicated a dearth in assessment strategies for evaluating Collaborator competencies in a surgical setting. An Interprofessional Collaborator Assessment Rubric was adopted in order to assess performance of Collaborator competencies through direct observation by orthopaedic preceptors. Ten staff surgeons assessed six residents on 25 competencies, using a four point Likert scale in both clinical and operative settings. The evaluations were collected and assessed for inter-rater reliability using Fleiss Kappa and percent agreement. **Results:** Weighted percent agreement was 80.6 percent and the mean Fleiss Kappa score was -0.218 (95% CI -0.406 to -0.085) demonstrating low inter-rater reliability. **Conclusion:** Despite the use of a validated assessment tool to evaluate the CanMEDs Collaborator role, inter-rater reliability results suggest low levels of assessor agreement. This project provides a framework for further assessments of collaborative competencies. There will be an increase in demand for evaluations of evaluation methods in the changing scope of medical education.

Appendix G: Medical Expert Role Abstract

Background: Orthopaedic surgical education in Canada has seen major change in the last 15 years. Work hour restrictions along with external influence have led to new approaches towards surgical training. With a change towards competency-based educational models under the CanMEDs headings there becomes a need to ensure the validity of modern evaluation methods.

Objective: To evaluate the reliability of a currently utilized surgical skill evaluation tool within an orthopaedic surgery residency program as measured by the Surgical Encounters Form.

Methods: A surgical evaluation tool has previously been created at our institution comprising 15 items spanning four of the CanMEDs competencies. Over a five-month period eleven staff members evaluated nine residents. Results were blinded to the primary investigator and coded by a third party. Eighty-eight evaluations were completed in total. The evaluations were collected and measurements of percent agreement, Cronbach's alpha, and Fleiss Kappa were obtained.

Results: Weighted percent agreement was 90.9 percent. Cronbach's alpha averaged 0.865, 0.920, 0.934, 1.00 and 1.00 for the Medical expert, Technical skills, Communicator, Collaborator, and Advocate roles respectively. The mean Fleiss Kappa score was 0.147 (95% CI -0.071 to 0.364) that demonstrated low inter-rater reliability. **Conclusion:** Despite the development of a validated assessment tool to evaluate surgical skills acquisition inter-rater reliability results suggest low levels of agreement between assessors.